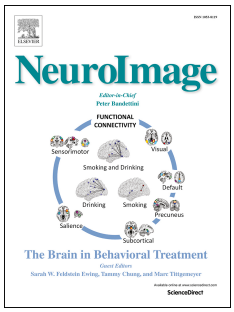
Accepted Manuscript



Spatially informed voxelwise modeling for naturalistic fMRI experiments

Emin Çelik, Salman Ul Hassan Dar, Özgür Yılmaz, Ümit Keleş, Tolga Çukura

PII: S1053-8119(18)32125-6

DOI: https://doi.org/10.1016/j.neuroimage.2018.11.044

Reference: YNIMG 15449

To appear in: *NeuroImage*

Received Date: 23 October 2018

Accepted Date: 25 November 2018

Please cite this article as: Çelik, E., Hassan Dar, S.U., Yılmaz, Öü., Keleş, Ü., Çukura, T., Spatially informed voxelwise modeling for naturalistic fMRI experiments, *NeuroImage* (2018), doi: https://doi.org/10.1016/j.neuroimage.2018.11.044.

Spatially Informed Voxelwise Modeling for Naturalistic fMRI Experiments

Emin Çelik [a,b], Salman Ul Hassan Dar[b,c], Özgür Yılmaz[b,c], Ümit Keleş[b,d], Tolga Çukur[a,b,c]

[a]Neuroscience Program, Bilkent University, Ankara, TR-06800, Turkey
[b]National Magnetic Resonance Research Center (UMRAM), Bilkent University, Ankara, TR-06800, Turkey
[c]Department of Electrical and Electronics Engineering, Bilkent University, Ankara, TR-06800, Turkey
[d]Division of Humanities and Social Sciences, California Institute of Technology, Pasadena, CA-91125, USA

Correspondence should be addressed to:
Tolga Çukur
Department of Electrical and Electronics Engineering
Bilkent University
Ankara, TR-06800, Turkey
E-mail: <cukur@ee.bilkent.edu.tr>

Running title: **Spatially Informed Voxelwise Modeling**
Manuscript summary:
Pages: 36
Abstract: 140 words
Introduction: 1053 words
Materials and Methods: 5464 words
Results: 2717 words
Discussion: 1574 words
Figures: 10
Supp. Figures: 18
Supp. Tables: 12

Abstract

Voxelwise modeling (VM) is a powerful framework to predict single voxel responses evoked by a rich set of stimulus features present in complex natural stimuli. However, because VM disregards correlations across neighboring voxels, its sensitivity in detecting functional selectivity can be diminished in the presence of high levels of measurement noise. Here, we introduce spatially-informed voxelwise modeling (SPIN-VM) to take advantage of response correlations in spatial neighborhoods of voxels. To optimally utilize shared information, SPIN-VM performs regularization across spatial neighborhoods in addition to model features, while still generating single-voxel response predictions. We demonstrated the performance of SPIN-VM on a rich dataset from a natural vision experiment. Compared to VM, SPIN-VM yields higher prediction accuracies and better capture locally congruent information representations across cortex. These results suggest that SPIN-VM offers improved performance in predicting single-voxel responses and recovering coherent information representations.

Keywords: fMRI, Voxelwise modeling, Response correlations, Coherent representation, Spatial regularization, Computational neuroscience

1. Introduction

Neural response correlations exist in multiple spatial scales across cortex, ranging from cortical columns with hundreds of neurons (Erwin et al., 1995) to neighborhoods of voxels in functional magnetic resonance imaging (fMRI) studies with hundreds of thousands of neurons (Zarahn et al., 1997). It is commonly hypothesized that these correlations reflect clustering of neural populations that form modules with specific functional selectivities, which leads to efficient information processing and coherent representation of information across cortex (Pouget et al., 2000; Schneidman et al., 2006). Consistent with this hypothesis, many fMRI studies have reported similar functional selectivity across neighboring voxels, suggesting coherent information representations. For instance, vision studies have shown that angle and eccentricity values are represented topographically in early visual areas (Engel et al., 1997; Tootell et al., 1998), and semantic information about object and action categories is represented in smooth gradients across higher-level visual areas and non-visual cortex (Huth et al., 2012).

The existence of spatial correlations in blood oxygen level dependent (BOLD) responses have often motivated traditional univariate analyses to perform spatial smoothing as a preprocessing step to improve signal-to-noise ratio (SNR). In the statistical parametric mapping (SPM) approach (Friston et al., 1994), spatial smoothing enables the use of Random Field Theory (Adler and Firman, 1981) to locate clusters with similar functional selectivity. In the functional localizer approach, spatial smoothing is used to locate a spatially contiguous set of voxels that are functionally selective to a certain stimulus class, such as faces (Kanwisher et al., 1997) or body parts (Downing et al., 2001). Traditional univariate analyses typically assume that functional selectivity is distributed homogeneously across neighborhoods, thereby ignoring differences in selectivity across individual voxels. As a consequence, the sensitivity to fine-grained information present in single voxels is reduced (Kriegeskorte and Bandettini, 2007).

An alternative approach that does not require explicit spatial smoothing is joint modeling of spatially contiguous voxels (Katanoda et al., 2002; Penny et al., 2005). In standard general linear modeling (GLM), a linear set of weights is estimated for each voxel that predicts the measured responses based on the stimulus or task timecourse (Nelder and Wedderburn, 1972). To improve sensitivity, the joint approach performs GLM on responses aggregated across a spatial neighborhood of voxels. One method is to estimate the model for the central voxel by uniformly weighing data from all voxels within the neighborhood (Katanoda et al., 2002). This uniform weighing renders joint modeling equivalent to spatial smoothing with a boxcar function across the neighborhood, and the interpretation of resulting models is difficult. A more recent method instead penalizes differences in model weights of voxels within the neighborhood (Penny et al., 2005). Because this previous method only employs spatial regularization, it can yield suboptimal sensitivity in the presence of a large number of model features or limited amount of measurements. This can be particularly limiting in the analysis of BOLD responses elicited by thousands of stimulus features during naturalistic experiments.

Another popular approach that avoids spatial smoothing is multivariate pattern analysis (MVPA). Building direct decoding models, MVPA analyzes the responses of multiple voxels in order to classify BOLD response patterns into a limited number of discrete experimental conditions (Haxby, 2012; Norman et al., 2006). While MVPA does not use spatial smoothing, classifier weights may not accurately reflect the contribution of individual voxels to the represented information because they are estimated for multiple voxels at once to optimize classification performance (Haufe et al., 2014). For example, a common MVPA method named searchlight analysis assumes that information is represented in small, localized clusters of voxels (Kriegeskorte et al., 2006). In searchlight analysis, a voxel at the center of a searchlight volume can be thought to represent significant stimulus information, merely because the volume contains other highly informative voxels (Etzel et

al., 2013). Thus, similar to joint modeling approaches (Katanoda et al., 2002; Penny et al., 2005), MVPA can be suboptimal in revealing information representations in single voxels.

In contrast to traditional fMRI analyses, voxelwise modeling (VM) is a powerful framework that offers improved sensitivity for fine-grained assessment of cortical representations in naturalistic fMRI experiments (Kay et al., 2008; Naselaris et al., 2009). Previous studies have demonstrated the elevated sensitivity of VM in examining the representations of diverse stimulus features in single voxels across cortex (Çukur et al., 2013b; Huth et al., 2012; Lescroart et al., 2015; Nishimoto et al., 2011). The goal of the VM framework is to assess functional selectivity at the finest resolution available—single voxels—in fMRI data. To do this, VM first constructs a model in the form of a dictionary of stimulus features (e.g., a set of object categories or a bank of spatiotemporal Gabor wavelets) that are hypothesized to elicit BOLD responses. For each voxel, VM then estimates the linearly-weighted combination of model features that best explain the measured BOLD responses (Naselaris et al., 2011). The model weights for each voxel reflect its selectivity to hundreds to thousands of model features that occur in natural stimuli. Note that VM employs regularization across model weights to prevent over-fitting to nuisance response variations. To increase sensitivity to single voxels, regularization parameters are optimized separately for each voxel using a cross-validation procedure performed on unsmoothed single-voxel responses. Once models are trained, model performance is evaluated on independent test data to ensure model generalizability. Because VM models each voxel independently without spatial smoothing, it enhances sensitivity for detecting functional selectivity in single voxels compared to traditional techniques (Dumoulin and Wandell, 2008; Mitchell et al., 2008; Serences and Saproo, 2012; Thirion et al., 2006). However, VM disregards potentially correlated information across neighboring voxels, yielding suboptimal sensitivity in the presence of high levels of measurement noise.

Here, we introduce spatially informed voxelwise modeling (SPIN-VM) to better utilize response correlations in neighboring voxels. To spatially inform the single-voxel models without smoothing, we utilize a weighted graph Laplacian based on inter-voxel distances (Grosenick et al., 2013; Penny et al., 2005). SPIN-VM performs regularization across both model features and spatial neighborhoods. While SPIN-VM enforces similar model weights across neighboring voxels, it still generates predictions of BOLD responses in single voxels. Therefore, it maintains high sensitivity to selectivity differences across individual voxels. We demonstrate SPIN-VM on an fMRI dataset collected in a natural vision experiment. Models obtained using VM and SPIN-VM are compared in terms of single-voxel prediction accuracy and local coherence of functional selectivity.

2. Materials and Methods

In this section, we first describe the experimental paradigm, data preprocessing and visualization techniques. We then introduce spatially informed voxelwise modeling (SPIN-VM) and explain its relationship to regular voxelwise modeling (VM). Finally, we describe local coherence analyses, and how effects of spatial smoothing were investigated.

2.1. Subjects

Five healthy male human subjects volunteered to participate in the study: S1 (age 25), S2 (age 25), S3 (age 25), S4 (age 32), and S5 (age 29). Participants had normal or corrected-to-normal vision. The experimental protocols were approved by the Institutional Review Board at the University of California, Berkeley (UCB). All participants gave written informed consent prior to scanning.

2.2. MRI acquisition parameters

Functional and anatomical MRI data were collected using a 3T Siemens Tim Trio scanner with a 32-channel head coil at the University of California, Berkeley. A gradient-echo echo-planar imaging (GE-EPI) sequence (TR=2 s, TE=34 ms, flip angle=74°, voxel size=2.24×2.24×3.5 mm^3, field-of-view=224×224 mm^2, 32 axial slices covering the entire cortex) was used to acquire T_2^*-weighted functional data. To avoid contamination from fat signal, the sequence was customized with a water-excitation radiofrequency (RF) pulse. Anatomical data were collected using a T_1-weighted magnetization-prepared rapid-acquisition gradient-echo (MP-RAGE) sequence (TR=2.30 s, TE=3.45 ms, flip angle=10°, voxel size=1×1×1 mm^3, field-of-view=256×256×192 mm^3). The anatomical data were used in order to reconstruct cortical surfaces for each subject. For two subjects, anatomical and retinotopic mapping data were collected using a 1.5T Philips Eclipse scanner.

2.3. Main experiment

The main experiment was conducted in three separate sessions. Color natural movies were shown to subjects and whole-brain BOLD responses were recorded in each session. Movies were selected from a diverse set of sources in order to avoid potential biases. High-definition movie frames were cropped to a square aspect ratio and downsampled to 512×512 pixels subtending 24°×24°. Participants were instructed to fixate on a centrally located color dot (0.16×0.16°) superimposed onto the movies. For continuous visibility, the color of the fixation dot changed at 3 Hz. An MRI-compatible projector (Avotec), a custom-built mirror system, and custom-designed presentation scripts were used for stimulus presentation. A total of 12 training runs and 9 testing runs were acquired across the three sessions. Different sets of movies were used for training and test runs, and the presentation order was interleaved in each session. Each training run contained 10 min of natural movies compiled by concatenating distinct 10-20 s movie clips without repetition. Each testing run contained 10 separate 1 min blocks in random order. Each block was presented nine times across three sessions and acquired BOLD responses were averaged across these repeats. Data collected during the first 10 s of each run were not used to minimize the effects of hemodynamic transients. These three sessions resulted in 3600 data samples for training and 270 data samples for testing. Note that these same data were analyzed in several recent studies (Çukur et al., 2016, 2013b, 2013a; Huth et al., 2012).

2.4. Functional localizers

Functional localizer data were acquired in two separate sessions. Category-selective brain areas were localized using six 4.5 min runs of 16 blocks, each lasting 16 s. Twenty static images were presented in each block, randomly selected from one of the following categories: objects, scenes, faces, body parts, animals, and spatially scrambled objects (Spiridon et al., 2006). The presentation order was randomized across runs. Each image was shown for 300 ms, followed by a 500 ms blank screen. Participants performed a one-back task to ensure they

maintained their focus on the experiment. Retinotopic areas were localized using four 9 min runs containing clockwise rotating polar wedges, counter-clockwise rotating polar wedges, expanding rings, and contracting rings (Hansen et al., 2007). Intraparietal sulcus was localized using one 10 min run of 30 blocks, each lasting 20 s and containing either a self-generated saccade task (among a pattern of targets) or a resting task (Connolly et al., 2000). Human motion processing complex (MT) was localized using four 90 s runs of 6 blocks, each lasting 15 s and containing either continuous or temporally scrambled natural movies (Tootell et al., 1995). Auditory cortex was localized in a single 10 min run consisting of 10 repeats of a 1 min auditory stimulus, which consisted of 20 s segments of speech, music, and natural sounds. Motor localizer data were collected in a single 10 min run during which subjects were cued to perform six different motor actions ("hand", "foot", "mouth", "saccade", "speech", "rest") in 20 s blocks in a random order.

2.5. Data preprocessing

FMRIB's Linear Image Registration Tool (FLIRT) (Jenkinson et al., 2002) was used for motion correction and image realignment. For each subject, functional brain volumes were aligned to the first image from the first functional run. Functional brain volumes were refined by removing non-brain tissue using Brain Extraction Tool (BET) (Smith, 2002). Low-frequency drifts in BOLD responses of individual voxels were removed using a median filter over a 120 s temporal window for each run, separately. The resulting time courses were z-scored individually for each voxel such that mean response across time points was 0 and standard deviation across time points was 1 for each voxel. No temporal or spatial smoothing was applied to the functional data from the main experiment. Motion correction and image realignment procedures were also applied to the functional localizer data such that the volumes are aligned to the first functional run from the main experiment. The localizer data were smoothed with a Gaussian kernel of full-width at half-maximum equal to 4 mm.

2.6. Visualization on cortical flatmaps

Cortical surfaces were reconstructed from T_1-weighted anatomical scans using FreeSurfer (Dale et al., 1999), separately for each hemisphere of each subject. After gray-white matter segmentation, five relaxation cuts were applied on the surface of each hemisphere and the surface crossing the corpus callosum was removed. Finally, the surfaces were flattened. Functional data were aligned to the anatomical data automatically using the FLIRT boundary-based alignment tool in the FSL library (Greve and Fischl, 2009). A six degree-of-freedom affine transformation was used in the three-dimensional voxel space. Registration accuracy was taken as the alignment error between the white-matter boundaries of the functional and anatomical data. For this procedure, the parameter "bbrtype" was set to "signed". Pycortex was used for surface projection (Gao et al., 2015). The resulting flatmaps were used for data visualization. Note that positive prediction scores indicate that a fit model explains meaningful variance in measured BOLD responses, whereas negative prediction scores indicate that the model does not explain any meaningful variance. Because model weights in a voxel with a negative prediction score do not accurately reflect its functional selectivity, it would be misleading to interpret the model weights. To prevent contamination from poorly modeled voxels, values of interest (e.g., category coefficients for cortical maps of semantic representation) were thresholded and scaled using a sigmoid function based on prediction scores for each voxel. This ensured that the lower the prediction score, the closer toward gray the color of the voxel moved (baseline gray level is 102 for Figs. 4, 6, and 7; and 51 for Figs. 8 and 9, range=0-255). As a result, model weights for voxels with positive prediction scores were visualized on cortical flatmaps, whereas model weights for voxels with negative prediction scores were masked. Note also that for all analyses reported in the manuscript, voxels were selected from the volumetric brain space. Cortical surfaces were used solely for visualization purposes.

2.7. Encoding models

2.7.1. Motion-energy model

We used a motion-energy model consisting of 2139 spatiotemporal Gabor filters to infer selectivities of single voxels for low-level visual features. The same motion-energy model was previously shown to accurately predict BOLD responses to natural movies in retinotopically organized early visual areas (Nishimoto et al., 2011). Each of the 2139 filters was a three-dimensional spatiotemporal sinusoid multiplied by a spatiotemporal Gaussian envelope. Filters were computed at six spatial frequencies (0, 1.5, 3, 6, 12, and 24 cycles/image), three temporal frequencies (0, 2, and 4 Hz), and eight directions (0, 45, 90, 135, 180, 225, 270, and 315°). Filters were positioned on a square grid that spanned 24°×24°. Filters at each spatial frequency were placed on the grid such that adjacent filters were separated by a distance of four standard deviations of the spatial Gaussian envelope. Then, to reduce dimensionality and improve model fits, a principal components analysis (PCA) was applied to the stimulus matrix. The first 400 PCs that explain 95.7% of the variance in the stimulus were selected.

2.7.1.1. Representation of low-level visual features

PCA was used to recover a group Gabor space from the motion-energy model weights of all subjects. Only voxels with highest prediction scores (top 10,000 for both VM and SPIN-VM) for each subject were included in estimating the group Gabor space to ensure high quality. Then, individual-subject model weights were projected onto the first three PCs of the group Gabor space for each cortical voxel to enable comparison of cortical representation across subjects. Subsequently, each voxel was assigned a color from RGB color space such that Gabor coefficients obtained by model weight projections in first, second, and third PCs represent red, green, and blue channels, respectively (see Fig. 7). We followed the in-silico simulation procedure outlined in (Nishimoto et al., 2011) to estimate selectivity for spatial frequency and eccentricity from the motion-energy model. In this procedure, the responses of each voxel to a two-dimensional dynamic Gaussian white noise pattern, presented at various positions across the virtual display, were estimated based on model weights. These predicted responses explain the sensitivity of each voxel to each position in space. As a result, each voxel was assigned discrete spatial frequency and eccentricity values based on motion-energy model weights. Similar colors imply selectivity for similar low-level visual features (e.g., magenta implies selectivity for low eccentricity and high spatial frequency). We identified four different colors that broadly correspond to distinct combinations of selectivity for spatial frequency and eccentricity.

2.7.2. Category model

We used a category model to infer selectivities of single voxels for distinct object and action categories present in the natural movie stimulus. The same category model was previously shown to accurately predict BOLD responses in high-level visual cortex (Çukur et al., 2013b; Huth et al., 2012). Object and action categories present in each 1 s portion of the movies were labeled using the WordNet lexicon (Miller, 1995). Superordinate categories entailed by each labeled category were also added to the list of features (i.e., categories) in accordance with the WordNet hierarchy. For example, if a portion of the movie was labeled with "car", it would also be labeled with "machine". After adding superordinate categories, a feature list with 1705 distinct object and action categories was formed. Time courses for all model features were obtained by aggregating the present/absent labels across the stimulus (see Fig. 1). Temporal downsampling was then applied to each time course to match the fMRI sampling rate. Then, to reduce dimensionality and improve model fits, a PCA was applied to the stimulus matrix. The first 300 PCs that explain 95.7% of the variance in the stimulus were selected. To minimize spurious correlations between global motion-energy and visual categories, a nuisance regressor was included that reflected the total motion energy in the movie stimulus. The total motion energy was obtained by summing the output of all spatiotemporal Gabor filters used in the motion-energy model.

2.7.2.1. Representation of semantic categories

PCA was used to recover a group semantic space from the category model weights of all subjects. Only voxels with highest prediction scores (top 10,000 for both VM and SPIN-VM) for each subject were included in estimating the group semantic space to ensure high quality. The first PC was observed to be highly correlated with the motion-energy in the movie stimulus (Huth et al., 2012), and therefore we did not use it when visualizing semantic representation across cortical surface. Due to the limitations of fMRI and a finite stimulus set, only the first few PCs will approximate the true underlying semantic space (Huth et al., 2012). Accordingly, individual-subject model weights were projected onto second, third, and fourth PCs of the group semantic space for each cortical voxel to enable comparison of cortical representation across subjects. Subsequently, each voxel was assigned a color from RGB color space such that category coefficients obtained by model weight projections in second, third, and fourth PCs represent red, green, and blue channels, respectively (see Fig. 6). Similar colors imply selectivity for similar semantic categories (e.g., dark blue implies selectivity for buildings and furniture). Six sets of broad categories (vehicles, buildings and furniture, animals, text and groups, humans and body parts, and geography) are identified with six different colors in Fig. 6 as examples. There are many other categories represented across cortex, these six categories are chosen only for visualization purposes.

2.8. Model fitting - VM

To fit category and motion-energy models, a voxelwise modeling framework was used (Çukur et al., 2013b; Huth et al., 2012). VM performs L_2-regularized linear regression to find model weights that describe how each model feature (e.g., object and action categories) influences measured BOLD responses (see Fig. 1). A category model was fit to measure tuning for high-level object and action categories (Huth et al., 2012). A separate motion-energy model was fit to measure tuning for elementary visual features such as spatiotemporal frequency and orientation (Nishimoto et al., 2011). To account for hemodynamic delays in BOLD responses, separate finite-impulse-response (FIR) filters were appended to each model feature. Temporal delays of two, three, and four samples (equivalently 4, 6, and 8 s) were applied by the FIR filters. To maximize the quality of fits, FIR coefficients were fit together with the model weights:

$$\begin{aligned} \mathbf{X} \quad &\times \quad \mathbf{W} \quad = \mathbf{Y} \\ [\mathbf{x}_{d4}\ \mathbf{x}_{d6}\ \mathbf{x}_{d8}] &\times [\mathbf{w}_{d4}\ \mathbf{w}_{d6}\ \mathbf{w}_{d8}]^{T} = \mathbf{Y} \end{aligned} \quad \textbf{(1)}$$

where $\mathbf{Y}$ is the response matrix of size (time points × N_{vox}), $\mathbf{X}$ is a stimulus matrix of size (time points × (3 × N_{feat})), and $\mathbf{W}$ is a matrix of size ((3 × N_{feat})× N_{vox}) that represents selectivity for model features, where N_{vox} is the number of voxels and N_{feat} is the number of model features. The subscripts d4, d6, and d8 denote the entries for each hemodynamic delay. Final selectivities were computed by averaging over delays.

VM estimates model weights via ridge regression

$$\min_{\mathbf{w}_i} \sum_i \|\mathbf{X}\mathbf{w}_i - \mathbf{y}_i\|_2^2 + \lambda_{feat} \sum_i \|\mathbf{w}_i\|_2^2 \ , i = 1, \dots, N_{vox} \quad \textbf{(2)}$$

where $\mathbf{w}_i$ is a vector of model weights, and $\mathbf{y}_i$ is a vector of BOLD responses for voxel i. The optimization problem in Eq. 2 is solved separately for each individual voxel. Eq. 2 is first compactly expressed in matrix form as

$$\min_{\mathbf{W}} \mathrm{Tr}\left(\mathbf{W}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{W}\right) + \lambda_{\mathrm{feat}}\mathrm{Tr}\left(\mathbf{W}\mathbf{W}^{\mathrm{T}}\right) - 2\mathrm{Tr}\left(\mathbf{W}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{Y}\right) + \mathrm{Tr}(\mathbf{Y}\mathbf{Y}^{\mathbf{T}}) \quad \textbf{(3)}$$

Minimization can then be performed by setting the gradient of the objective with respect to **W** to zero

$$(\mathbf{K} + \lambda_{\mathrm{feat}}\mathbf{I})\mathbf{W} = \mathbf{M} \quad \textbf{(4)}$$

where $\mathbf{K} = \mathbf{X}^{\mathrm{T}}\mathbf{X}$, and $\mathbf{M} = \mathbf{X}^{\mathrm{T}}\mathbf{Y}$. **K** reflects the auto-covariance of model features and **M** reflects the cross-covariance of model features and BOLD responses. Finally, the solution to Eq. 4 can be obtained by a pseudoinverse operation

$$\mathbf{W}^{*} = (\mathbf{K} + \lambda_{\mathrm{feat}}\mathbf{I})^{\dagger}\mathbf{M} \quad \textbf{(5)}$$

A 10-fold cross-validation procedure was used to optimize the regularization parameter across features (λ_{feat}) for each voxel, and the regularization parameter resulting in the highest prediction score across cross-validation folds was selected. In each fold, 10% of the training data were randomly held out, with the remaining 90% being used to fit models. Prediction score was taken as the correlation coefficient (Pearson's r) between the measured and predicted BOLD responses. Raw correlation coefficients are biased downward by noise in the measured BOLD responses (David and Gallant, 2005). Therefore, correlation coefficients were corrected for noise bias following the procedure detailed in (Huth et al., 2012). Finally, models were refit to the entire training data using optimal regularization parameters in a single step. Note that all model fitting, evaluation, and comparisons were done based on voxelwise model fits in individual subjects.

A 1,000-fold jackknife resampling (at a rate of 80%) procedure was used to calculate prediction scores on independent test data in order to assess model performance. Average prediction score across jackknife iterations was calculated. Custom software written in Matlab (MathWorks) was used for all model fitting procedures. Mean prediction scores across ROIs were also calculated for each subject independently. Then, a single mean prediction score (± std) was calculated for each ROI via bootstrapping across subjects. By performing our calculations in each subject's individual brain space and not transforming every subject's data onto a common anatomical space, we avoided any bias or distortion that could contaminate the results. On the other hand, we averaged ROI-wise prediction scores from each subject to draw broader inferences about statistical comparison of competing methods, following common procedure in voxelwise analyses (see Sprague and Serences, 2013). To test for significant differences between two competing methods, prediction scores were randomly sampled with replacement across subjects, and the mean difference between the methods was computed. To determine the significance level, 10,000 bootstrap samples were generated, and p-value was taken as the fraction of bootstrap samples where the mean difference is less than 0 (for right-sided tests) or greater than 0 (for left-sided tests). An identical sampling procedure was used to assess the significance of differences in local coherence values.

In addition, to test whether using L1-norm could be a viable alternative to dimensionality reduction based on PCA in conjunction with L2-norm regularization across model features, we fit voxelwise models using L1-norm (without applying PCA) and calculated prediction scores. For this analysis, we used 14 regularization parameters spanning the range $[2^{-2}, 2^{11}]$ for both the category and motion-energy models. A coordinate descent algorithm was employed to solve the L1 minimization problem. We found that the PCA-based approach yields significantly higher prediction scores than the L1-norm approach across all functional ROIs ($p < 0.05$; Supp.

Fig. 16). For instance, mean prediction scores for FFA were (0.7146 ± 0.0318) and (0.5243 ± 0.0556) for the PCA-based approach and the L1-norm approach, respectively. Similarly, mean prediction scores for the whole cortex were (0.1735 ± 0.0114) and (0.1014 ± 0.0051) for the PCA-based approach and the L1-norm approach, respectively. This finding is in line with a previous study from our laboratory that reports that L2-norm regularization of model weights yields superior performance to L1-norm regularization in FFA (Çukur et al., 2013a). As a result, we did not consider L1-norm thereafter.

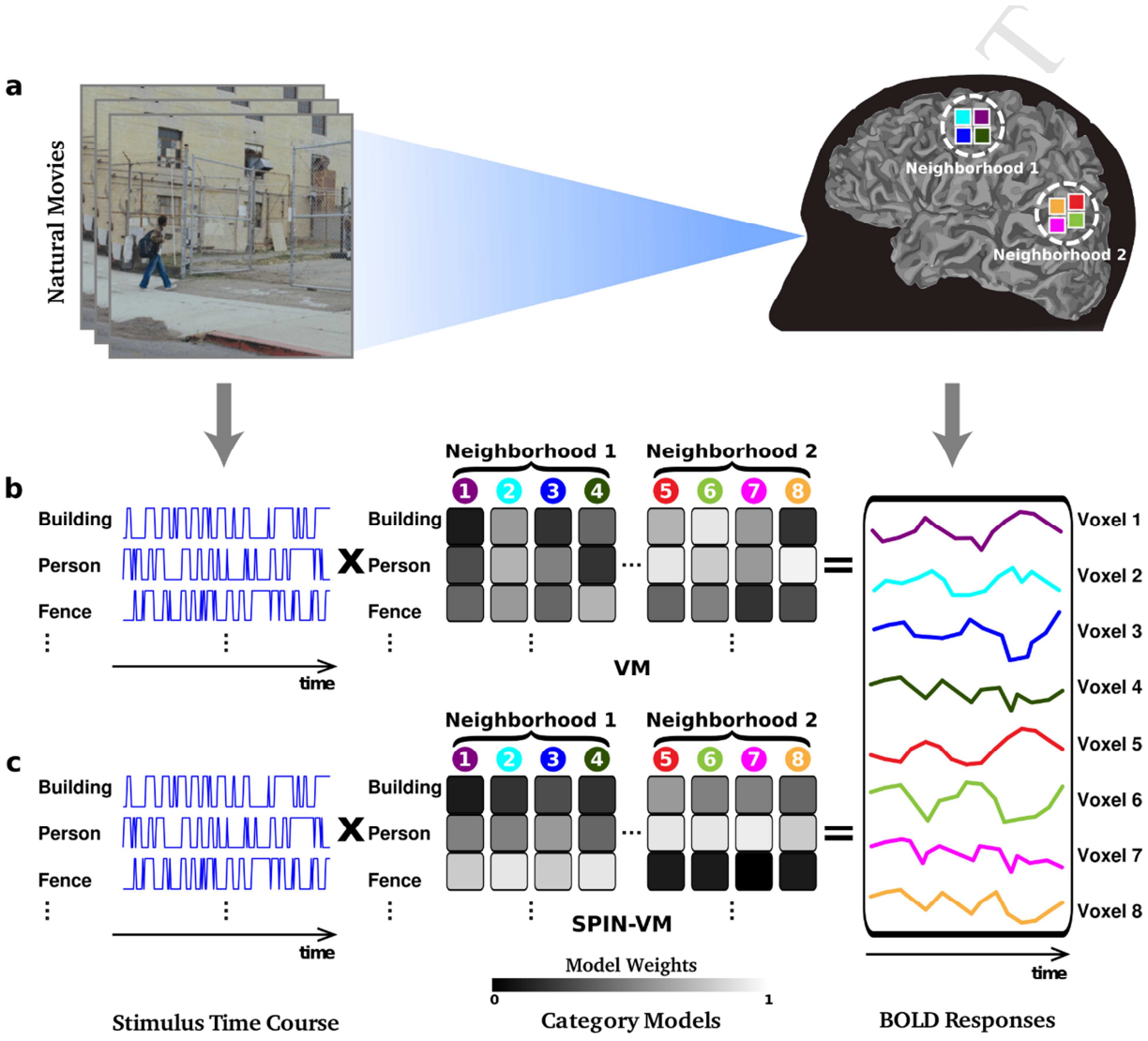


Fig. 1. Experimental paradigm and model fitting. (a) Subjects viewed natural movies and whole-brain BOLD responses were recorded using fMRI. Functional selectivity in single voxels was estimated in individual subjects using voxelwise modeling (VM) and spatially informed voxelwise modeling (SPIN-VM). Model fitting procedures for VM and SPIN-VM are illustrated here for a category model. Model weights reflect the selectivity of individual voxels for 1705 distinct object and action categories. **(b)** In VM, each voxel is modeled independently from its neighbors. High levels of noise in measured BOLD responses can cause nuisance variability in estimated model weights (model weights for two distinct neighborhoods of voxels illustrated). **(c)** In SPIN-VM, each voxel is modeled while utilizing shared information across neighborhoods of voxels to enhance sensitivity during model fits. As a result, it can more accurately assess functional selectivity in single voxels even in the presence of high levels of noise.

2.9. Model fitting – SPIN-VM

VM has been shown to produce powerful and informative results in fine-grained assessment of cortical

representations (Çukur et al., 2016, 2013b, 2013a, Huth et al., 2016, 2012; Nishimoto et al., 2011). Since the VM framework does not perform any spatial smoothing across voxels or subjects, it enhances sensitivity for detecting selectivity in single voxels compared to standard analysis techniques such as SPM (Friston et al., 1994). However, in the presence of high levels of measurement noise, VM may yield suboptimal sensitivity as it disregards correlated responses across neighboring voxels. To leverage shared information across neighboring voxels, SPIN-VM implements regularization not only across the feature dimension as in VM, but also across neighborhoods of voxels. To obtain optimal solutions, we enforce constraints on both rows and columns of the unknown weight matrix (Subbian and Banerjee, 2013). Utilizing shared information across voxels naturally increases estimation sensitivity of model weights and it also prevents unnecessary smoothing across the feature dimension to beat noise.

In SPIN-VM, a spatial regularization term is used to take into account spatial neighborhood information across voxels

$$\sum_{(i,j)\in N_{nei}} c_{ij}\left\|\mathbf{w}_i - \mathbf{w}_j\right\|_2^2 \tag{6}$$

where N_{nei} is the set of voxels in a neighborhood. By selecting an appropriate set of filter weights, c_{ij}, that are large for voxels in close proximity and small for voxels that are far apart from each other, this term enforces neighboring voxels to have relatively similar weights. The spatial regularization term is added to the original optimization problem for VM in Eq. 2. The objective function that leverages information across neighboring voxels then becomes

$$\min_{\mathbf{w}_i} \sum_i \|\mathbf{X}\mathbf{w}_i - \mathbf{y}_i\|_2^2 + \lambda_{feat}\sum_i \|\mathbf{w}_i\|_2^2 + \lambda_{nei}\sum_{(i,j)\in N_{nei}} c_{ij}\left\|\mathbf{w}_i - \mathbf{w}_j\right\|_2^2, i = 1, \ldots, N_{vox} \tag{7}$$

where the third term is the spatial regularizer and λ_{nei} is the corresponding regularization parameter. It can be shown that the spatial regularizer in Eq. 7 can be compactly expressed in terms of a graph Laplacian matrix $\mathbf{L}$ such that

$$\lambda_{nei}\sum_{(i,j)\in N_{nei}} c_{ij}\left\|\mathbf{w}_i - \mathbf{w}_j\right\|_2^2 = \lambda_{nei}\mathrm{Tr}(\mathbf{W}\mathbf{L}\mathbf{W}^T) \tag{8}$$

Following the transition from Eq. 2 to Eq. 3, the entire objective function can then be written as

$$\min_{\mathbf{W}} \mathrm{Tr}\left(\mathbf{W}^T\mathbf{X}^T\mathbf{X}\mathbf{W}\right) + \lambda_{feat}\mathrm{Tr}\left(\mathbf{W}\mathbf{W}^T\right) + \lambda_{nei}\mathrm{Tr}\left(\mathbf{W}\mathbf{L}\mathbf{W}^T\right) - 2\mathrm{Tr}\left(\mathbf{W}^T\mathbf{X}^T\mathbf{Y}\right) + \mathrm{Tr}(\mathbf{Y}\mathbf{Y}^T) \tag{9}$$

Finally, minimization can be achieved by setting the gradient of the objective with respect to $\mathbf{W}$ to zero

$$(\mathbf{K} + \lambda_{feat}\mathbf{I})\mathbf{W} + \lambda_{nei}\mathbf{W}\mathbf{L} = \mathbf{M} \tag{10}$$

The expression in Eq. 10 can be simplified by defining $\mathbf{A} = (\mathbf{K} + \lambda_{feat}\mathbf{I})$, and $\mathbf{B} = \lambda_{nei}\mathbf{L}$ such that $\mathbf{A}\mathbf{W} + \mathbf{W}\mathbf{B} = \mathbf{M}$. Regularization over rows of $\mathbf{W}$ is performed by $\mathbf{A}$, which reflects auto-covariance of model features. Regularization over columns of $\mathbf{W}$ is performed by $\mathbf{B}$, which is based on a graph Laplacian containing spatial

proximity information across neighborhoods of voxels. Unlike VM where Eq. 4 can be solved via a simple pseudoinverse, the solution of Eq. 10 in SPIN-VM requires a more elaborate algorithm outlined in *Pseudocode for SPIN-VM* below. In steps, the eigenvalue decomposition of $\mathbf{A}$ is calculated for each λ_{feat} separately and the eigenvalues $\mathbf{D}$, and the eigenvectors $\mathbf{Q}$ are stored. Schur decomposition of $\mathbf{L}$ is computed, where $\mathbf{L} = \mathbf{USU}^{\mathrm{T}}$, prior to solving Eq. 10. This enables an efficient solution because Schur decomposition of a symmetric matrix gives a diagonal matrix $\mathbf{S}$ that simplifies subsequent calculations. This decomposition is then used to calculate $\mathbf{P}$ for each λ_{nei} separately, such that $\mathbf{P} = \mathbf{1}./(\mathbf{D}_{\mathrm{r}} + \lambda_{\mathrm{nei}}\mathbf{S}_{\mathrm{r}})$, where $\mathbf{D}_{\mathrm{r}}$ is a matrix constructed by repeating $\mathbf{D}_{\mathrm{d}}$ and $\mathbf{S}_{\mathrm{r}}$ is a matrix constructed by repeating $\mathbf{S}_{\mathrm{d}}$ (see *Pseudocode for SPIN-VM* below). $\mathbf{D}_{\mathrm{d}}$ is a column vector that contains the diagonal elements of $\mathbf{D}$, $\mathbf{S}_{\mathrm{d}}$ is a row vector that contains the diagonal elements of $\mathbf{S}$, and $./$ denotes elementwise division. Finally, the solution for each (λ_{feat}, λ_{nei}) pair is obtained

$$\mathbf{W}^{*} = -\mathbf{Q}(\mathbf{P}.^{*}(\mathbf{Q}^{\mathrm{T}}\mathbf{MU}))\mathbf{U}^{\mathrm{T}} \quad \textbf{(11)}$$

where $.^{*}$ denotes elementwise multiplication. Here, $\mathbf{W}$ was separately estimated for each (λ_{feat}, λ_{nei}) pair. To compare the cortical distribution of regularization parameters between VM and SPIN-VM, we employed the same range of λ_{feat} for both techniques. λ_{nei} also spanned the same range as λ_{feat}. We used 10 regularization parameters spanning the range $[2^5, 2^{14}]$ for the category model, and 13 regularization parameters spanning the range $100 \times [2^5, 2^{17}]$ for the motion-energy model. The same 10-fold cross-validation procedure as in VM was used for SPIN-VM to select the optimal (λ_{feat}, λ_{nei}) pair independently during model fitting. The pair of regularization parameters that resulted in the highest prediction scores across cross-validation folds were recorded as optimal regularization parameters for each voxel separately. Models were refit using the (λ_{feat}, λ_{nei}) pair that gives the highest prediction scores (see *Model fitting - VM*). Prediction scores were assessed using the same jackknifing procedure as in VM.

Pseudocode for SPIN-VM

Input: **X:** stimulus matrix of size (time points) × ($3\times N_{feat}$)
Y: response matrix of size (time points) × (N_{vox})
K: auto-covariance matrix of size ($3\times N_{feat}$) × ($3\times N_{feat}$), where $\mathbf{K} = \mathbf{X}^{\mathrm{T}}\mathbf{X}$
M: cross-covariance matrix of size ($3\times N_{feat}$) × (N_{vox}), where $\mathbf{M} = \mathbf{X}^{\mathrm{T}}\mathbf{Y}$
λ_{feat}: regularization parameter for features
λ_{nei}: regularization parameter for neighbors
A: auto-covariance matrix of size ($3\times N_{feat}$) × ($3\times N_{feat}$), where $\mathbf{A} = (\mathbf{K} + \lambda_{feat}\mathbf{I})$
L: Laplacian matrix of size (N_{vox}) × (N_{vox}), which stores proximity information between voxels
B: Laplacian matrix of size (N_{vox}) × (N_{vox}), where ($\mathbf{B} = \lambda_{nei}\mathbf{L}$)

Output: W: model weight matrix of size ($3\times N_{feat}$) × (N_{vox})

Precompute Schur decomposition of $\mathbf{L}$, such that $\mathbf{L} = \mathbf{USU}^{\mathrm{T}}$
Save $\mathbf{U}$ and $\mathbf{S}_{\mathrm{d}}$, where $\mathbf{S}_{\mathrm{d}} = \mathrm{diag}(\mathbf{S})$ is a row vector of size $1 \times (N_{vox})$

Solve: ($\mathbf{AW} + \mathbf{WB} = \mathbf{M}$)
begin
  for λ_{feat}:
    Find eigenvalues of $\mathbf{A}$ and store them in $\mathbf{D}$
    Set $\mathbf{D}_{\mathrm{d}} = \mathrm{diag}(\mathbf{D})$, where $\mathbf{D}_{\mathrm{d}}$ is a column vector of size ($3\times N_{feat}$) × 1
    Set $\mathbf{D}_{\mathrm{r}} = [\mathbf{D}_{\mathrm{d}}, \mathbf{D}_{\mathrm{d}}, \cdots]$ where $\mathbf{D}_{\mathrm{d}}$ repeats N_{vox} times such that $\mathbf{D}_{\mathrm{r}}$ is of size ($3\times N_{feat}$) × (N_{vox})
    Find eigenvectors of $\mathbf{A}$ and store them in $\mathbf{Q}$
    for λ_{nei}:

Set $\mathbf{S}_r = \begin{bmatrix} \mathbf{S}_d \\ \mathbf{S}_d \\ \vdots \end{bmatrix}$ where $\mathbf{S}_d$ repeats $(3\times N_{feat})$ times such that $\mathbf{S}_r$ is of size $(3\times N_{feat}) \times (N_{vox})$

$\mathbf{P} = \mathbf{1}./(\mathbf{D}_r + \lambda_{nei}\mathbf{S}_r)$, where $\mathbf{P}$ is of size $(3\times N_{feat}) \times (N_{vox})$

$\mathbf{W}^* = -\mathbf{Q}(\mathbf{P}.^*(\mathbf{Q}^T\mathbf{M}\mathbf{U}))\mathbf{U}^T$

2.9.1. Hyperparameters

The hyperparameters of SPIN-VM include the regularization parameters λ_{feat} and λ_{nei}. In addition, there are two hyperparameters that shape the Laplacian matrix: window size and filter type. The Laplacian matrix $\mathbf{L}$ is of size $N_{vox} \times N_{vox}$, where N_{vox} is the number of cortical voxels. $\mathbf{L} = \mathbf{T} - \mathbf{C}$, where c_{ij} (entries of matrix $\mathbf{C}$) corresponds to the proximity of voxels i and j in three-dimensional space (high for immediate neighbors, low or zero for voxels far away), and $\mathbf{T}$ is a diagonal matrix with $\mathbf{T}_{ii} = \sum_j c_{ij}$. Both window size and filter type determine c_{ij}.

Window size relates to the selection of voxel neighborhoods across which spatial regularization is performed. One possibility is to select voxels that are in close spatial proximity to each other ("spatial neighborhood"); another possibility is to select voxels that are functionally similar to each other ("functional neighborhood"). We investigated both. To optimize the extent of spatial neighborhood for SPIN-VM, we tested seven different window sizes (extending 3, 5, 7, 9, 11, 13, or 15 voxels). Note that a window size of 3 voxels is the smallest size we can test without breaking symmetry as it indicates only a single voxel on each side of the central voxel whereby a neighborhood of 27 voxels is constructed. For example, a window size of 1 would simply indicate a single voxel with no neighbors—a case that is equivalent to VM. Only voxels within the specified window were considered neighbors, and thus included in the construction of the graph Laplacian. Specifically, a cubic window was prescribed in which zero weights were assigned to voxels outside the window. When part of the cube was outside the cortex, the voxels outside were assigned zero weights regardless of their proximity to the central voxel. We set the filter type to Gaussian for this analysis whereby selected voxels in the neighborhood were weighted based on a Gaussian function. Separate Laplacian matrices based on a Gaussian filter were formed using each window size. We determined the optimal window size by comparing prediction scores across functional ROIs (see Supp. Tables 1 and 2).

Similarly, to optimize the extent of functional neighborhood for SPIN-VM, we tested the same seven window sizes (extending 3, 5, 7, 9, 11, 13, or 15 voxels). To measure functional similarity between voxels, we computed pairwise correlations in BOLD responses. The functional neighborhood of each voxel was formed from voxels that show the highest correlations with the given voxel (e.g., 125 voxels for window size 5). Similar to spatial neighborhood analysis, we set the filter type to Gaussian for this analysis whereby selected voxels in the neighborhood were weighted based on a Gaussian function. Separate Laplacian matrices based on a Gaussian filter were formed using each window size. We determined the optimal window size for functional neighborhoods by comparing prediction scores across functional ROIs.

The primary difference between spatial and functional neighborhood analyses is the difference in calculation of inter-voxel distances, according to which a set of neighboring voxels is selected. For spatial neighborhoods, selection is based on Euclidean distance between voxels in three-dimensional volumetric space. For functional neighborhoods, selection is based on (1-R), where R is the correlation coefficient between response vectors of voxels.

In this study, filter type determines the distribution of entries of $\mathbf{C}$ within a specified neighborhood. We tested three types of filters: Gaussian filter, average (or boxcar) filter, and LoG (Laplacian of Gaussian) filter. As an

alternative, we also investigated the case where weights are assigned based on functional correlations between voxel responses rather than the abovementioned filters. The Gaussian filter was centered on the voxel of interest and had a FWHM equal to half the window size. The tails of the Gaussian function stretched towards the edges of the cube and dropped to zero outside the edges:

$$c_{ij} = \frac{\exp(-(|x_i - x_j|^2 + |y_i - y_j|^2 + |z_i - z_j|^2)/(2\sigma^2))}{\sum_{(i,j)\in N_{nei}} \exp(-(|x_i - x_j|^2 + |y_i - y_j|^2 + |z_i - z_j|^2)/(2\sigma^2))} \quad \textbf{(12)}$$

where x_i, y_i, z_i are the coordinates of voxel i in the three-dimensional grid of voxels. The average filter assigned uniform weights to all voxels in the neighborhood such that the sum of weights equaled 1:

$$c_{ij} = \frac{1}{|N_{nei}|} \quad \textbf{(13)}$$

where N_{nei} is the set of cortical voxels in the neighborhood. The LoG filter was a rotationally symmetric filter with identical standard deviation to the Gaussian filter:

$$c_{ij} = \frac{\exp(-(|x_i - x_j|^2 + |y_i - y_j|^2 + |z_i - z_j|^2)/(2\sigma^2)) \cdot (|x_i - x_j|^2 + |y_i - y_j|^2 + |z_i - z_j|^2 - 2\sigma^2)}{\sigma^4 \sum_{(i,j)\in N_{nei}} \exp(-(|x_i - x_j|^2 + |y_i - y_j|^2 + |z_i - z_j|^2)/(2\sigma^2))} \quad \textbf{(14)}$$

We determined the optimal filter type by comparing prediction scores across functional ROIs (see Supp. Tables 3 and 4).

2.10. Effects of spatial smoothing

In VM, shared information across neighboring voxels is ignored, therefore VM might have suboptimal sensitivity in assessment of functional selectivity. To increase sensitivity in the presence of high levels of noise, one alternative approach would be to smooth BOLD responses prior to model fitting. While smoothing may help reduce noise by averaging across multiple voxels, it can decrease sensitivity in detecting selectivity in single voxels. Thus, it can lead to undesirable loss of spatial precision (Kamitani and Sawahata, 2010). In contrast, SPIN-VM uses spatial regularization to leverage shared information across neighboring voxels without any averaging. SPIN-VM still estimates model weights for individual voxels and generates predictions for raw unsmoothed single-voxel BOLD responses. Therefore, SPIN-VM improves model performance while maintaining sensitivity in detecting functional selectivity in individual voxels.

To test the effects of spatial smoothing, we performed response smoothing via a centered Gaussian low-pass filter of size 3×3×3 with FWHM equal to half the window size, the same filter that was used to form graph Laplacians. We then implemented the standard model fitting procedures as in VM on these smoothed BOLD responses. We calculated prediction scores and local coherence values for both the category and motion-energy models. To demonstrate that SPIN-VM is fundamentally different than smooth-VM, we compared the prediction scores and local coherence values of models obtained using these two different procedures. Note that while training and validation takes place on smoothed responses, prediction scores are still calculated on unsmoothed responses for smooth-VM. For SPIN-VM, however, no smoothing was applied on training, validation, or test data.

Smoothing inherently suppresses nuisance variations in BOLD responses including physiological and

measurement noise. As a result, smoothing test data is likely to cause an upward bias in prediction score measurements. To examine this issue, we first measured the prediction scores of models obtained via VM, smooth-VM and SPIN-VM on smoothed test data. Training and validation data for VM and SPIN-VM were unsmoothed, and training and validation data for smooth-VM were smoothed for this analysis. Furthermore, we measured the prediction scores of models obtained via VM, smooth-VM and SPIN-VM when both test and validation data were smoothed. Training data for VM and SPIN-VM were unsmoothed, and training data for smooth-VM were smoothed for this analysis.

2.11. Effect of training data size

Since SPIN-VM uses spatial information across multiple voxels unlike VM, we expected that it would yield higher prediction performance for single voxels compared to VM. This improved performance can be particularly valuable when the size of training data is limited. To investigate this issue, we fit separate models using both VM and SPIN-VM using training samples of three different sizes; we used the full set (3600 samples), a half set (1800 samples), and a quarter set (900 samples). For each size, model prediction scores were calculated on the independent test data. Percentage improvement compared to VM for all functional ROIs was calculated as

$$\text{improvement}(\%) = \frac{r_x - r_{VM}}{1 - \min(r_x, r_{VM})} \times 100 \quad \textbf{(15)}$$

where r_x denotes the mean prediction score obtained by method x, where x is either SPIN-VM or smooth-VM and r_{VM} denotes the mean prediction score obtained by VM. This measure normalizes the raw improvement against the maximum possible improvement. Note that this measure is bias-free as it is possible to obtain a negative "improvement" in cases where r_{VM} is larger than r_x.

2.12. Local coherence analysis

It is commonly thought that the human brain encodes similar information across spatially clustered groups of neural populations (Pouget et al., 2000). Studies on low-level vision suggest that retinotopic features such as spatiotemporal frequency and orientation are represented coherently in early-visual areas (Tootell et al., 1998). A recent study further suggests that semantic information is represented in smooth gradients across much of cerebral cortex (Huth et al., 2012). These previous studies imply that neighboring cortical voxels typically represent correlated information. If such correlation exists, the implication is that these voxels have similar feature selectivity and thus they should have coherent model weights. Because VM fits an independent model to each voxel, it might be less sensitive in capturing this coherence. SPIN-VM, on the other hand, explicitly leverages correlated information rendering it more sensitive in capturing coherent functional selectivity. Thus, we expect that selectivity maps obtained via SPIN-VM will be more coherent on the cortical surface compared to those obtained via VM.

To test this prediction, we computed local coherence values for each cortical voxel. Spatial variability of each model feature was taken as the standard deviation of feature weights across voxels in a 3×3×3 neighborhood:

$$\sigma_{\text{weights}} = \sum_{i=1}^{F} \sqrt{\frac{1}{N-1} \sum_{j=1}^{N} \left| \mathbf{w}_{ij} - \frac{1}{N} \sum_{j=1}^{N} \mathbf{w}_{ij} \right|^2} \quad \textbf{(16)}$$

where N is the number of cortical voxels in the neighborhood, F is the number of features retained after PCA

(300 for the category model, 400 for the motion-energy model), and $\mathbf{w}_{ij}$ is the selectivity of voxel j for feature i. The spatial variability values given by VM, SPIN-VM, and smooth-VM were then normalized by the maximum value across the three methods, and then inverted to obtain local coherence values. We calculated local coherence of an ROI by averaging across voxels within the ROI.

3. Results

SPIN-VM utilizes three additional hyperparameters during model fitting compared to VM. The first one is λ_{nei}, the spatial regularization parameter across neighborhoods of voxels. λ_{nei} is selected for each voxel independently during model fitting to control the relative degree of regularization in feature vs. spatial dimensions. The other two are window size and filter type, which determine the characteristics of spatial regularization. Although neighboring cortical voxels typically represent correlated information, the extent and distribution of these correlations can vary across cortical areas. To account for potential variability, we performed spatial regularization by utilizing a graph Laplacian matrix that stores proximity information among voxels. To keep the number of variables to a minimum, we selected the optimal window size and filter type through a rigorous optimization procedure and used these optimal parameters thereafter.

3.1. Parameter optimization for SPIN-VM

An important concern for SPIN-VM is the selection of voxel neighborhoods. One possibility is to select voxels that are in close spatial proximity ("spatial neighborhood"); another possibility is to select voxels that are functionally similar ("functional neighborhood"). We investigated both possibilities. To optimize the extent of spatial neighborhood for SPIN-VM, we tested seven different window sizes (extending 3, 5, 7, 9, 11, 13, or 15 voxels). Two different encoding models were used. The first one was a category model that measured selectivity for object and action categories. The second one was a motion-energy model that measured selectivity for low-level visual features including spatiotemporal frequency and orientation. We fit separate category and motion-energy models independently for each window size. Prediction scores across well-known functional ROIs are listed in Supp. Table 1 for the category model, and in Supp. Table 2 for the motion-energy model. A window size of 3 yields the highest prediction scores for the category model across the majority of the ROIs ($p < 0.05$, Bootstrap test). For the motion-energy model, a window size of 3 yields the highest prediction scores across all ROIs ($p < 0.05$, Bootstrap test).

Similarly, to optimize the extent of functional neighborhood for SPIN-VM, we tested the same seven window sizes (3, 5, 7, 9, 11, 13, or 15 voxels). To measure functional similarity between voxels, we computed pairwise correlations in BOLD responses. The functional neighborhood of each voxel was formed from voxels that show the highest correlations with the given voxel (e.g., 125 voxels for window size 5). When these functional neighborhoods are used, a window size of 9 yields the highest prediction scores for the category model across the majority of the ROIs ($p < 0.05$; see Supp. Fig. 14 that shows the improvement in prediction scores with a window size of 9 over a window size of 15). However, prediction scores based on spatial neighborhoods are still higher than those based on functional neighborhoods for the category model across the majority of the ROIs (56.2% of voxels across ROIs prefer a spatial window of 3 over a functional window of 9, the remaining voxels have similar prediction scores for both cases; Supp. Fig. 11). Thus, we used a spatial neighborhood with a window size of 3 voxels for subsequent analyses to ensure high model performance and low model complexity.

Another important design parameter for SPIN-VM is how information from neighboring voxels is weighted. Within a given neighborhood, it is expected that correlations among neurons will diminish with increasing distance (Lee et al., 1998; Smith and Kohn, 2008). However, the precise dependence between response correlation and spatial distance is unknown. In SPIN-VM, responses from neighboring voxels are used to improve the accuracy of the central voxel's model. To optimize the relative weighting of these responses we tested three different types of filters: Gaussian, average, and Laplacian of Gaussian (LoG). As an alternative, we also investigated the case where weights are assigned based on functional correlations between voxel responses rather than the abovementioned filters. We fit separate category and motion-energy models independently for each filter type. Prediction scores across well-known functional ROIs are listed in Supp. Table 3 for the category

model, and in Supp. Table 4 for the motion-energy model. Gaussian filter yields the highest prediction scores for both the category and motion-energy models across the majority of the ROIs (45.6% and 42.1% of voxels across ROIs prefer Gaussian filter for the category and motion-energy models, respectively. The remaining voxels have similar prediction scores for all filter types). Gaussian filter also yields higher prediction scores for the category model across the majority of the ROIs compared to the alternative approach of using functional correlations between voxel responses (52.2% of voxels across ROIs prefer Gaussian filters over functional correlations, the remaining voxels have similar prediction scores for both cases; Supp. Fig. 12). Based on these results, we determined the optimal hyperparameters to be a Gaussian filter with a window size of 3 for both the category and motion-energy models.

3.2. Prediction performance of SPIN-VM

Because SPIN-VM utilizes correlated information across neighboring voxels, we expect that it will improve model performance compared to VM. To examine this issue, we fit separate category and motion-energy models in single voxels using VM and SPIN-VM. Prediction scores obtained using the full set of training data for each functional ROI are listed in Supp. Table 5 for the category model and in Supp. Table 6 for the motion-energy model. We calculated the improvement in prediction scores ("SPIN-VM vs. VM" and "SPIN-VM vs. smooth-VM") across twelve ROIs for the category and motion-energy models (Figs. 2 and 3, respectively). For both models, SPIN-VM outperforms VM in all ROIs ($p < 0.05$, Bootstrap test). For the category model, improvements up to 10% are observed in high-level visual areas across lateral occipitotemporal cortex and ventral temporal cortex, including FFA, EBA, PPA, MT, and LOC. For the motion-energy model, the improvements are relatively more uniform (up to 7%) across early- and high-level visual areas.

We further expect that these improvements in prediction accuracy will grow as the size of the training data becomes limited. With fewer training data, models are likely to overfit and thus poorly generalize to test data. Since SPIN-VM utilizes shared information across neighboring voxels, it can alleviate the performance loss that VM and smooth-VM can experience. To test this prediction, we fit separate models using the half set (1800 samples), and quarter set (900 samples) of training data. We calculated the improvement in prediction scores (SPIN-VM vs. VM) across twelve ROIs for both the category and motion-energy models based on each set (Fig. 2). With both half and quarter sets, SPIN-VM outperforms VM in all ROIs for the category and motion-energy models ($p < 0.05$). As expected, when moving from the full set to the quarter set, the improvements with SPIN-VM significantly increase (up to 17%) in the category-selective areas in ventral temporal cortex for the category model ($p < 0.05$). Similarly, the improvements with SPIN-VM significantly increase (up to 11%) in all ROIs for the motion-energy model ($p < 0.05$). Taken together, these results indicate that SPIN-VM improves the performance of single-voxel models, and that these improvements become more prominent for smaller sets of training data.

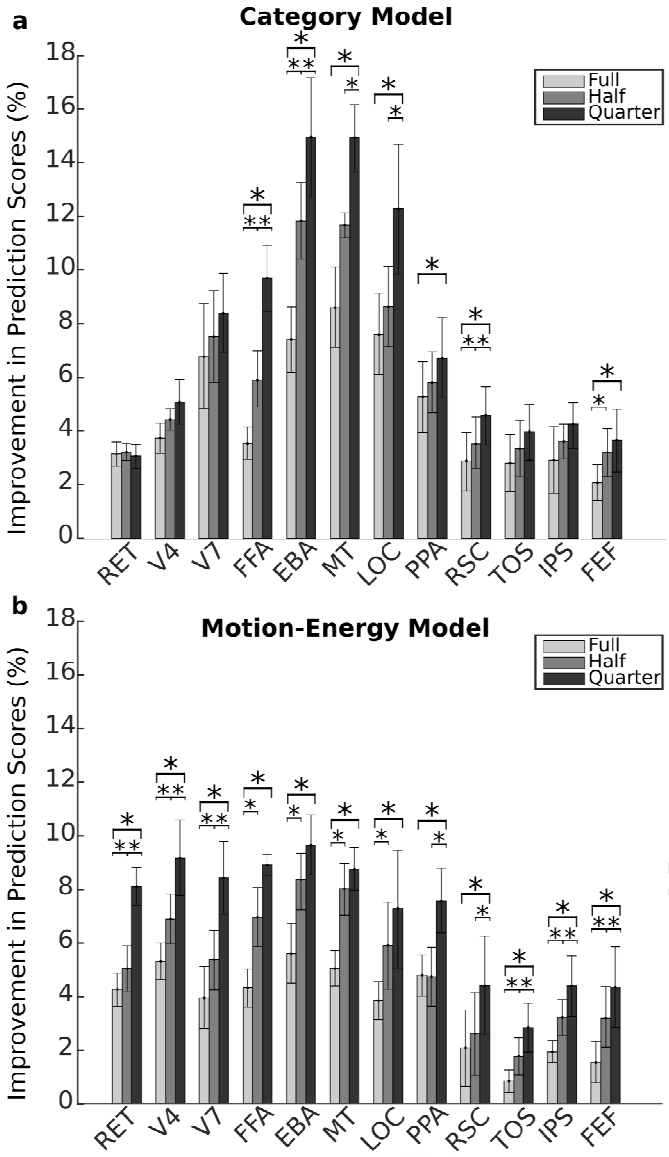


Fig. 2. Prediction score improvements with SPIN-VM over VM. Improvement in prediction scores with SPIN-VM over VM, displayed in twelve functional ROIs. Prediction scores were estimated separately while the size of training data was varied: Full set (light gray), half (gray), quarter (dark gray). Prediction scores are shown as mean percentage improvement across five subjects. Error bars indicate standard error of the mean (SEM). Brackets indicate significant differences across conditions corresponding to different sizes of training data ($p < 0.05$, Bootstrap test). **(a)** Improvements for the category model. **(b)** Improvements for the motion-energy model. For both models and regardless of the size of training data, SPIN-VM significantly improves prediction scores in all functional ROIs compared to VM ($p < 0.05$). For the category model, the largest improvements are observed in high-level visual areas across lateral occipitotemporal cortex and ventral temporal cortex, including FFA, EBA, PPA, MT, and LOC. As expected, these improvements become significantly larger as the size of training data is reduced ($p < 0.05$). For the motion-energy model, improvements in prediction scores are relatively more uniform across early- and high-level visual areas. Similar to the category model, the improvements in prediction scores with the motion-energy model become significantly larger as the size of training data is reduced ($p < 0.05$). Abbreviations: EBA, extrastriate body area; FEF, frontal eye fields; FFA, fusiform face area; IPS, intraparietal sulcus; LOC, lateral occipital complex; MT, human middle temporal area; PPA, parahippocampal place area; RET, early visual areas V1-3; RSC, retrosplenial cortex; TOS, transverse occipital sulcus.

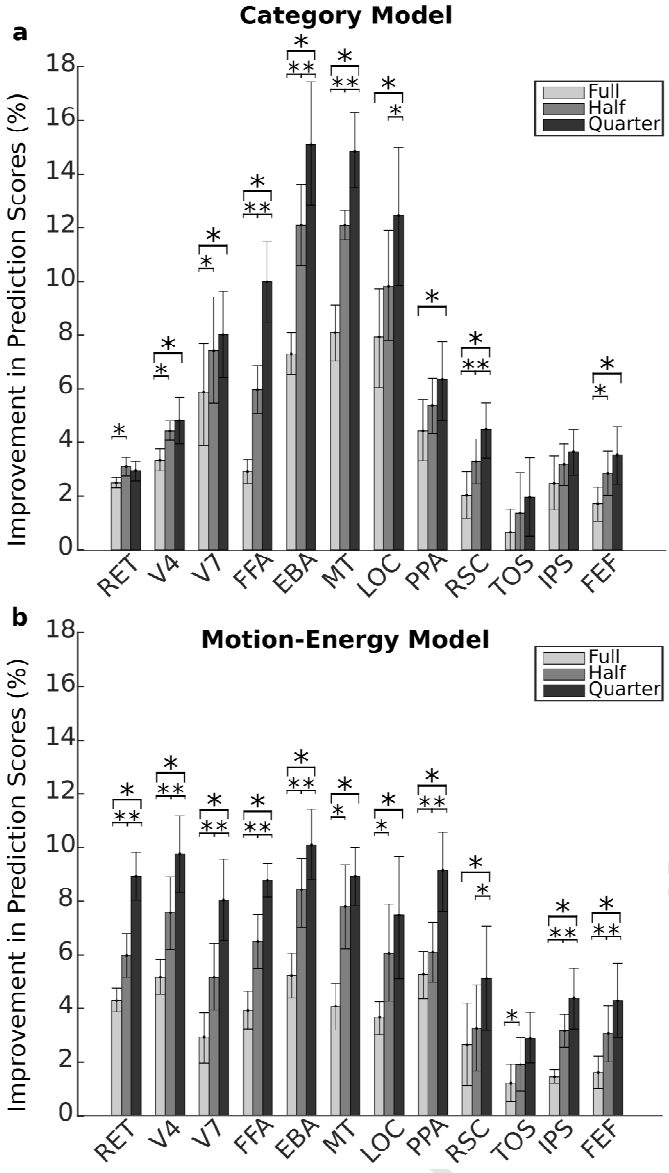


Fig. 3. Prediction score improvements with SPIN-VM over smooth-VM. Improvement in prediction scores with SPIN-VM over smooth-VM, displayed in twelve functional ROIs. Prediction scores were estimated separately while the size of training data was varied: Full set (light gray), half (gray), quarter (dark gray). Prediction scores are shown as mean percentage improvement across five subjects. Error bars indicate standard error of the mean (SEM). Brackets indicate significant differences across conditions corresponding to different sizes of training data ($p < 0.05$, Bootstrap test). **(a)** Improvements for the category model. **(b)** Improvements for the motion-energy model. For both models and regardless of the size of training data, SPIN-VM significantly improves prediction scores in all functional ROIs compared to smooth-VM ($p < 0.05$). The only exception is TOS, where SPIN-VM and smooth-VM perform similarly for the category model ($p = 0.1424$). For the category model, the largest improvements are observed in high-level visual areas across lateral occipitotemporal cortex and ventral temporal cortex, including FFA, EBA, PPA, MT, and LOC. As expected, these improvements become significantly larger as the size of training data is reduced ($p < 0.05$). For the motion-energy model, improvements in prediction scores are relatively more uniform across early- and high-level visual areas. Similar to the category model, the improvements in prediction scores with the motion-energy model become significantly larger as the size of training data is reduced ($p < 0.05$).

Broad improvements in prediction scores with SPIN-VM imply the existence of correlated information across many regions of cortex. In the presence of such correlations, one could argue that an alternative approach would be to apply a simple smoothing across BOLD responses prior to VM. Spatial smoothing can help alleviate measurement noise, however it will inadvertently decrease sensitivity to functional selectivity differences across voxels as it inherently suppresses nuisance variations in BOLD responses. An important advantage of spatial regularization over smoothing is that regularization parameters can be optimized separately for each voxel in each subject. An equally important advantage is that cross-validation procedures used to select regularization parameters (thereby model weights) and assess model performance can be performed on unsmoothed data, in order to retain maximal sensitivity to information represented in single voxels. In contrast, cross-validation on smoothed responses optimizes parameters and measures performance inherently for a population of voxels, and so it can yield suboptimal sensitivity to single voxels. Thus, we expect that SPIN-VM will outperform naive spatial smoothing in terms of model performance. To examine this issue, we calculated prediction score improvements with SPIN-VM over smooth-VM for three different sizes of training data (full, half, quarter) and for both the category and motion-energy models (Fig. 3). Note that smooth-VM was trained and validated on

smoothed BOLD responses but tested on unsmoothed responses. The resulting prediction scores are listed in Supp. Table 5 for the category model and in Supp. Table 6 for the motion-energy model. SPIN-VM performs significantly better in all ROIs for both the category and motion-energy models ($p < 0.05$). The only exception is TOS, where SPIN-VM and smooth-VM perform similarly for the category model ($p = 0.1424$). This result indicates that spatial regularization of model weights is more effective than spatial smoothing in utilizing shared information across neighboring voxels.

Next, we investigated the effect of testing on smoothed responses. We found that prediction scores for all three methods are elevated when smoothed test data were used, even though VM and SPIN-VM models were fit to and validated on unsmoothed data (Supp. Fig. 17, Supp. Tables 9-10). Compared to measurements on unsmoothed test data, mean prediction scores across whole cortex increase by 12% for VM, 15% for SPIN-VM, and 24% for smooth-VM (naturally smooth-VM benefits relatively more from smoothed test data). We also measured the prediction scores of models obtained via VM, smooth-VM and SPIN-VM when both test and validation data were smoothed. In this case, we find that SPIN-VM yields nearly identical performance to smooth-VM (Supp. Fig. 18, Supp. Tables 11-12). Taken together, these results suggest that higher prediction scores for voxelwise models measured on smoothed responses do not necessarily indicate improved model performance, but they can rather reflect a statistical bias.

3.3. Sensitivity in measuring selectivity for model features

VM performs regularization across model features during model fitting. As a result, heavier regularization parameters will be prescribed in the presence of high measurement noise, reducing sensitivity to inter-voxel selectivity differences. In contrast, the additional spatial regularization in SPIN-VM can help subdue unnecessary regularization across model features. Therefore, we expect that SPIN-VM will be more sensitive in detecting selectivity for distinct model features compared to VM.

To investigate this issue, we compared the optimal λ_{feat} values when using the category model for VM and SPIN-VM by visualizing them on cortical flatmaps. SPIN-VM exhibits more conservative regularization across model features compared to VM, especially across early- and high-level visual areas in occipital and ventral temporal cortices (Fig. 4). To illustrate the effect of λ_{feat} on estimated model weights, we illustrate functional selectivity differences between VM and SPIN-VM for a representative voxel in intraparietal sulcus (IPS) (Fig. 5). A substantially lower regularization parameter is used across features for this voxel with SPIN-VM ($\lambda_{feat} = 2^5$) compared to VM ($\lambda_{feat} = 2^{14}$). Importantly, the response of this voxel is well-estimated with SPIN-VM ($r = 0.73$), but not with VM ($r = 0.05$). IPS has been implicated in the representation of actions and locomotion of animate objects (Grefkes and Fink, 2005). While the model obtained via VM fails to capture selectivity for these features, SPIN-VM successfully captures selectivity for categories related to animals such as ‘rodent’ and ‘carnivore’, as well as categories related to movement such as ‘move’ and ‘jump’. This result suggests that SPIN-VM prevents unnecessary overpenalization across model features and improves sensitivity in estimating functional selectivity for individual features.

We also visualized the optimal λ_{nei} values on cortical flatmaps (Supp. Fig. 13). As expected, we find that optimal λ_{nei} values are relatively higher in both low-level retinotopic and high-level category selective visual areas that are more engaged during viewing of natural movies than non-visual areas such as frontotemporal, motor, and somatosensory cortices. These high λ_{nei} values likely compensate for the relatively lower λ_{feat} values in SPIN-VM compared to VM.

Finally, we inspected the functional selectivity profiles of individual voxels as measured by SPIN-VM and

smooth-VM. A representative voxel in posterior superior temporal sulcus (pSTS) is illustrated (Supp. Fig. 15; similar to Fig. 5). pSTS has been implicated in the representation of facial identities and visually observed social interactions (Srinivasan et al., 2016; Walbrin et al., 2018). While the model obtained via smooth-VM largely fails to capture these representations, SPIN-VM successfully captures selectivity for categories related to individuals such as ‘person’ and ‘man’, as well as categories related to social communication such as ‘talk’ and ‘text’. This simple example clearly demonstrates that smoothing reduces sensitivity to functional selectivity in individual voxels.

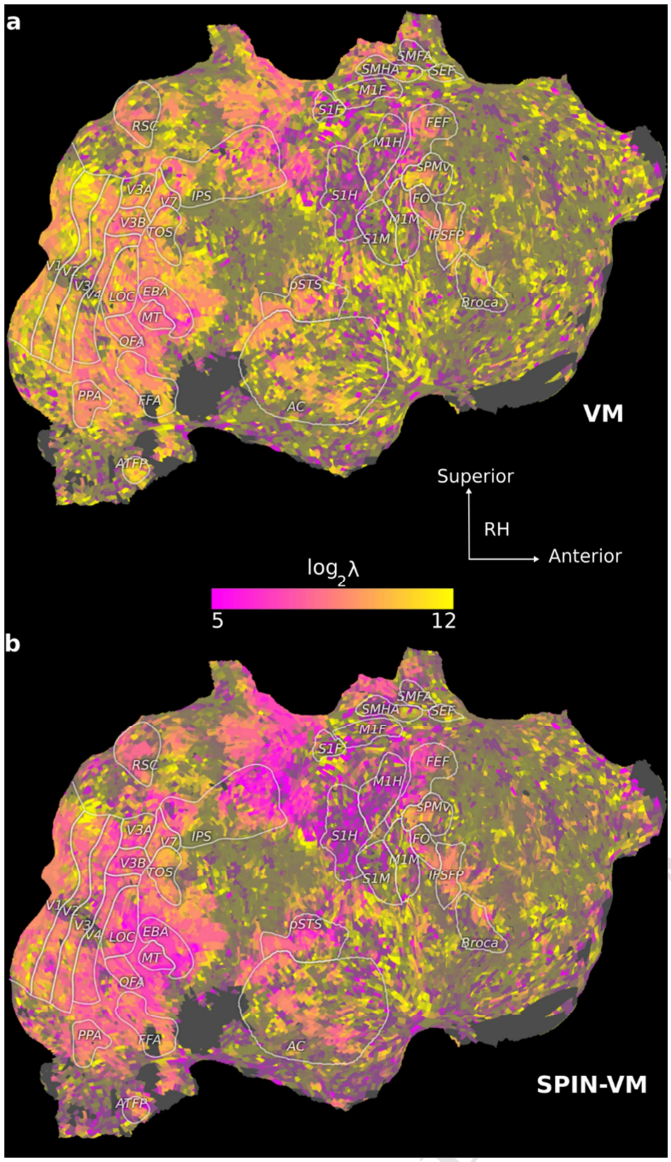


Fig. 4. Cortical distribution of regularization parameters. Cortical flatmaps of optimal regularization parameters across model features (λ_{feat}) for **(a)** VM and **(b)** SPIN-VM displayed in subject S1 for the category model. Optimal λ_{feat} values were determined separately for each voxel during model fitting. Color bar shows the range of λ_{feat} [2^5-2^{12}] in logarithmic scale (pink = low, yellow = high). Prescribing higher λ_{feat} enforces increased smoothing across the feature weights in the model. Therefore, it reduces sensitivity in capturing potential selectivity for distinct features. In contrast, prescribing lower λ_{feat} improves sensitivity. Optimal λ_{feat} values are much lower with SPIN-VM compared to VM, especially across early- and high-level visual areas in occipital and ventral temporal cortices. Therefore, SPIN-VM is more sensitive in capturing potential selectivity for individual features. White labels and outlines denote brain regions identified using conventional functional localizers. Dark gray denotes brain regions with fMRI signal dropout. RH, right hemisphere. AC, auditory cortex; ATFP, anterior temporal face patch; Broca, Broca's area; FO, frontal opercular eye movement area; IFSFP, inferior frontal sulcus face patch; M1F, M1H, M1M, primary motor areas for feet, hands, and mouth; OFA, occipital face area; S1F, S1H, S1M, primary somatosensory areas for feet, hands, and mouth; S2F, secondary somatosensory area for feet; SEF, supplementary eye fields; SMFA, supplementary motor foot area; SMHA, supplementary motor hand area; sPMv, superior premotor ventral speech area.

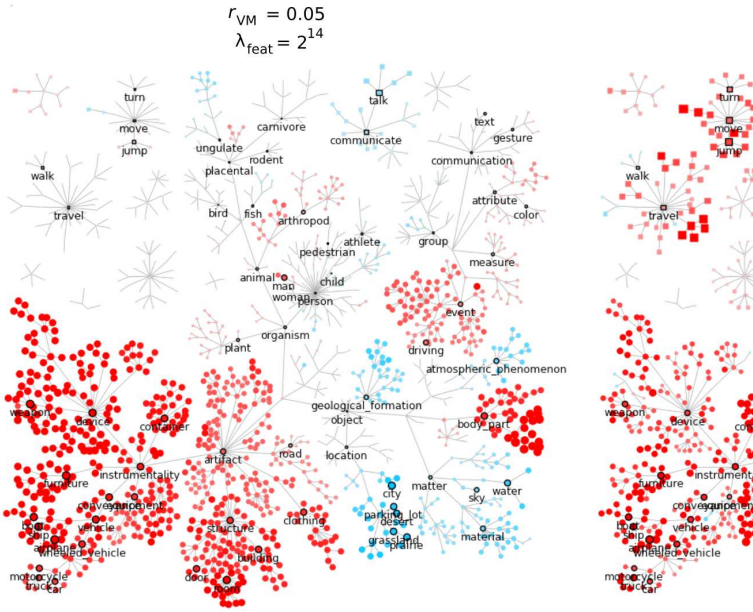


Fig. 5. Functional selectivity in a single voxel. Functional selectivity for object and action categories as measured by the category model for a single voxel (voxel #35890) in intraparietal sulcus (IPS) of subject S1. Functional selectivity obtained by VM (left) and SPIN-VM (right) is shown. Each node in these graphs represents a distinct object or action organized according to the hierarchical relations in the WordNet lexicon. Some important nodes are labeled to orient the reader. Red nodes correspond to categories that evoke above-mean responses, whereas blue nodes correspond to categories that evoke below-mean responses. The size of each node reflects the magnitude of the category response. The response of voxel #35890 is well-predicted by SPIN-VM (r = 0.73), and only poorly-predicted by VM (r = 0.05). Note that in VM, a

substantially larger regularization parameter (λ_{feat}) was used across features. This reduces sensitivity and predictive power of models obtained via VM. In contrast, SPIN-VM applies a relatively lenient regularization across features, and it has greater sensitivity in capturing selectivity for a broader distribution of categories. IPS has been implicated in the representation of actions and locomotion of animate beings (Grefkes and Fink, 2005). While the model obtained via VM fails to capture these representations, SPIN-VM successfully captures selectivity for categories related to animals such as 'rodent' and 'carnivore', as well as categories related to movement such as 'move' and 'jump'.

3.4. Local coherence of cortical representations

It is commonly assumed that the human brain encodes information coherently across spatially clustered groups of neural populations (Pouget et al., 2000). Consistent with this view, studies on low-level vision suggest that visual space is represented topographically in early-visual areas where nearby voxels represent similar angle and eccentricity values (Engel et al., 1997; Tootell et al., 1998). A recent study on natural vision further suggests that semantic information is also represented in smoothly organized gradients across much of cerebral cortex (Huth et al., 2012). These results indicate that both high-level category and low-level motion-energy representations in cortex exhibit a substantial degree of spatial coherence.

Because SPIN-VM explicitly leverages correlated information across neighboring voxels, it can offer increased sensitivity to unravel spatially coherent cortical representations compared to VM. To examine this issue, we compared the high-level category and low-level motion-energy representations recovered using VM and SPIN-VM. We first formed separate lower-dimensional spaces (a semantic space for the category model and a Gabor space for the motion-energy model) by applying PCA on fit model weights. We then projected individual-subject model weights onto these lower-dimensional spaces (see *Methods* for details). For a representative subject, Fig. 6 displays the semantic maps (see Supp. Figs. 1-5 for all subjects) and Fig. 7 displays the Gabor maps (see Supp. Figs. 6-10 for all subjects). The figures include cortical flatmaps of semantic and low-level visual representation based on model weights estimated using the full (left) and a quarter (right) set of the training data. We observe that SPIN-VM yields more coherent semantic and Gabor maps compared to VM. The difference between the two methods is clearer when only a quarter of the training data is used. The improved coherence in semantic maps is clearly seen across high-level visual areas in lateral occipitotemporal cortex and ventral temporal cortex that are implicated in semantic representation during natural vision (Huth et al., 2012). Similarly, the improved coherence in Gabor maps is particularly noticeable across early visual areas that are implicated in representation of low-level visual information (Engel et al., 1997; Tootell et al., 1998). Taken together, these results indicate that SPIN-VM is more powerful in recovering coherent representations compared to VM.

Next, a voxelwise metric was used to quantitatively evaluate the improvement in coherence of model weights. Spatial variability of each model feature was taken as the standard deviation of feature weights across a neighborhood and then inverted to obtain local coherence values. Local coherence was calculated in single voxels for both the category and motion-energy models. These coherence values were projected onto the cortical surface to illustrate differences in local coherence (category model, Fig. 8; motion-energy model, Fig. 9). Mean local coherences within functional ROIs were also calculated to draw statistical inferences on competing methods, the same procedure as the one used for calculating mean prediction scores across ROIs was employed (Fig. 10; listed in Supp. Table 7 for the category model; listed in Supp. Table 8 for the motion-energy model).

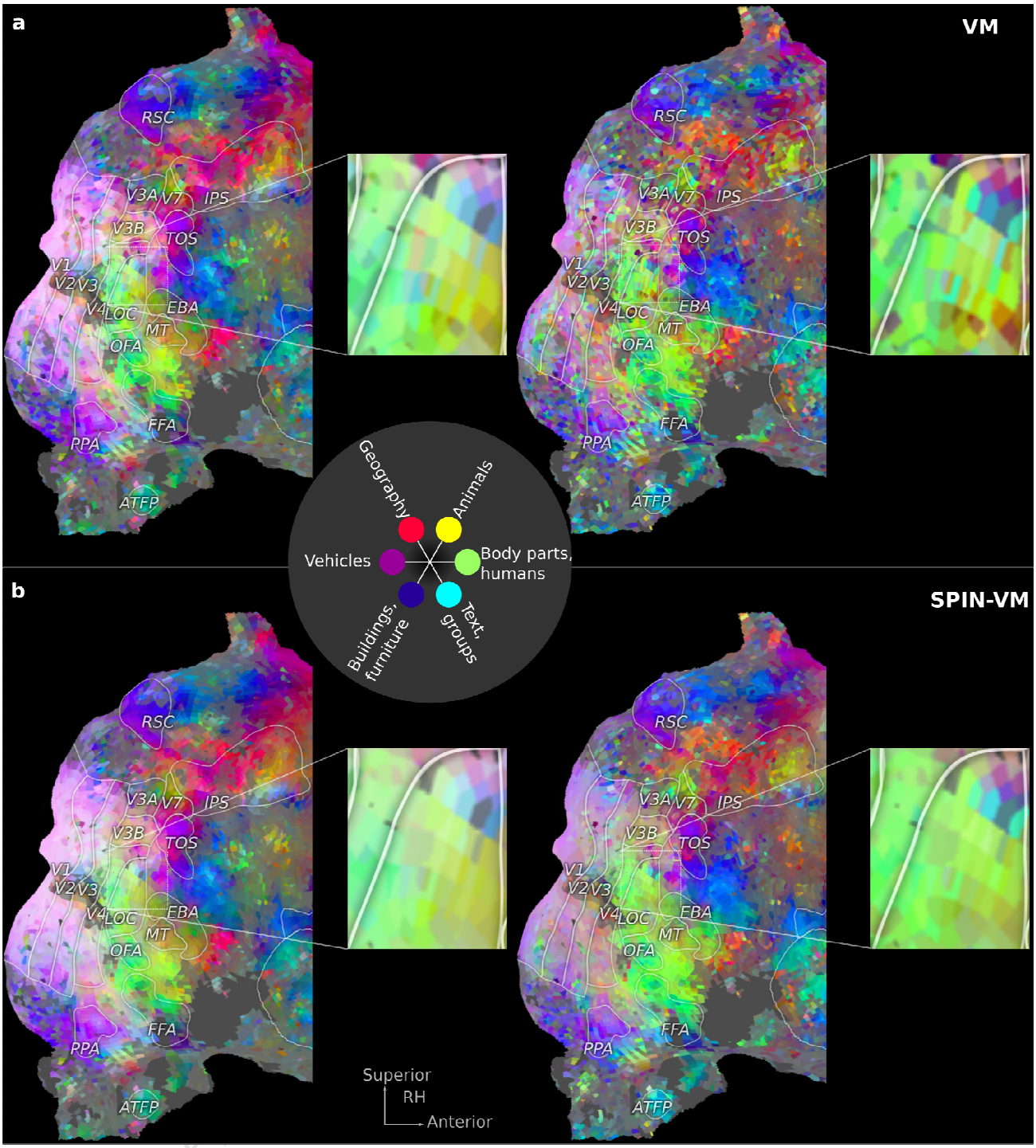


Fig. 6. Cortical flatmaps of semantic representation. Cortical flatmaps of semantic representation as measured by **(a)** VM and **(b)** SPIN-VM for subject S1. The flatmaps on the left are generated based on the model weights estimated using the full training data, whereas the flatmaps on the right are generated based on the model weights estimated using one quarter of the training data. To obtain consistent principal components (PCs) across both VM and SPIN-VM models, model weights obtained by both techniques were pooled and PCA was applied. Category model weights for each voxel were then projected onto the second, third, and fourth PCs of the group semantic space. Each voxel was assigned a color by representing projections on the second, third, and fourth PCs with red, green, and blue channels, respectively. Similar colors imply selectivity for similar semantic categories (e.g., dark blue implies selectivity for buildings and furniture, whereas magenta implies selectivity for vehicles). Insets show zoomed-in views of a cortical region in and around LOC. Compared to VM, estimated selectivities of neighboring voxels are more congruent (i.e., they have more similar colors) for

SPIN-VM regardless of whether models are trained on a full or a quarter set. The difference, however, is more pronounced when they are trained on a quarter set. Therefore, SPIN-VM produces more coherent semantic maps across many high-level visual areas. Formatting is identical to Fig. 4.

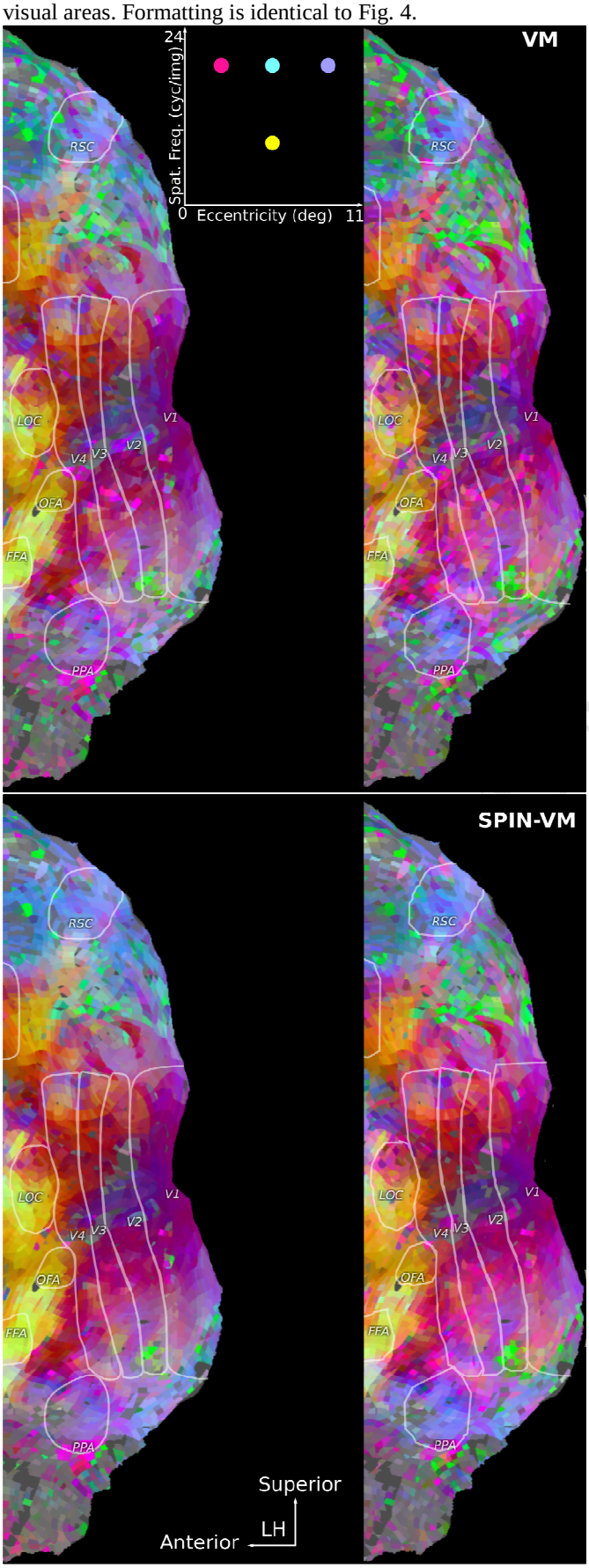


Fig. 7. Cortical flatmaps of low-level visual representation. Cortical flatmaps of low-level visual representation as measured by VM (top) and SPIN-VM (bottom) for subject S5. The flatmaps on the left are generated based on the model weights estimated using the full training data, whereas the flatmaps on the right are generated based on the model weights estimated using one quarter of the training data. To obtain consistent principal components (PCs) across both VM and SPIN-VM models, model weights obtained by both techniques were pooled and PCA was applied. Motion-energy model weights for each voxel were then projected onto the first three PCs of the group Gabor space. Each voxel was assigned a color by representing projections on the first, second, and third PCs with red, green, and blue channels, respectively. Similar colors imply selectivity for similar low-level properties (e.g., yellow signifies medium eccentricity and lower spatial frequency, whereas magenta signifies low eccentricity and higher spatial frequency). Compared to VM, estimated selectivities of neighboring voxels are more congruent (i.e., they have more similar colors) for SPIN-VM regardless of whether models are trained on a full or a quarter set. The difference, however, is more pronounced when they are trained on a quarter set. Therefore, SPIN-VM produces more coherent Gabor maps across early visual areas. Formatting is identical to Fig. 4.

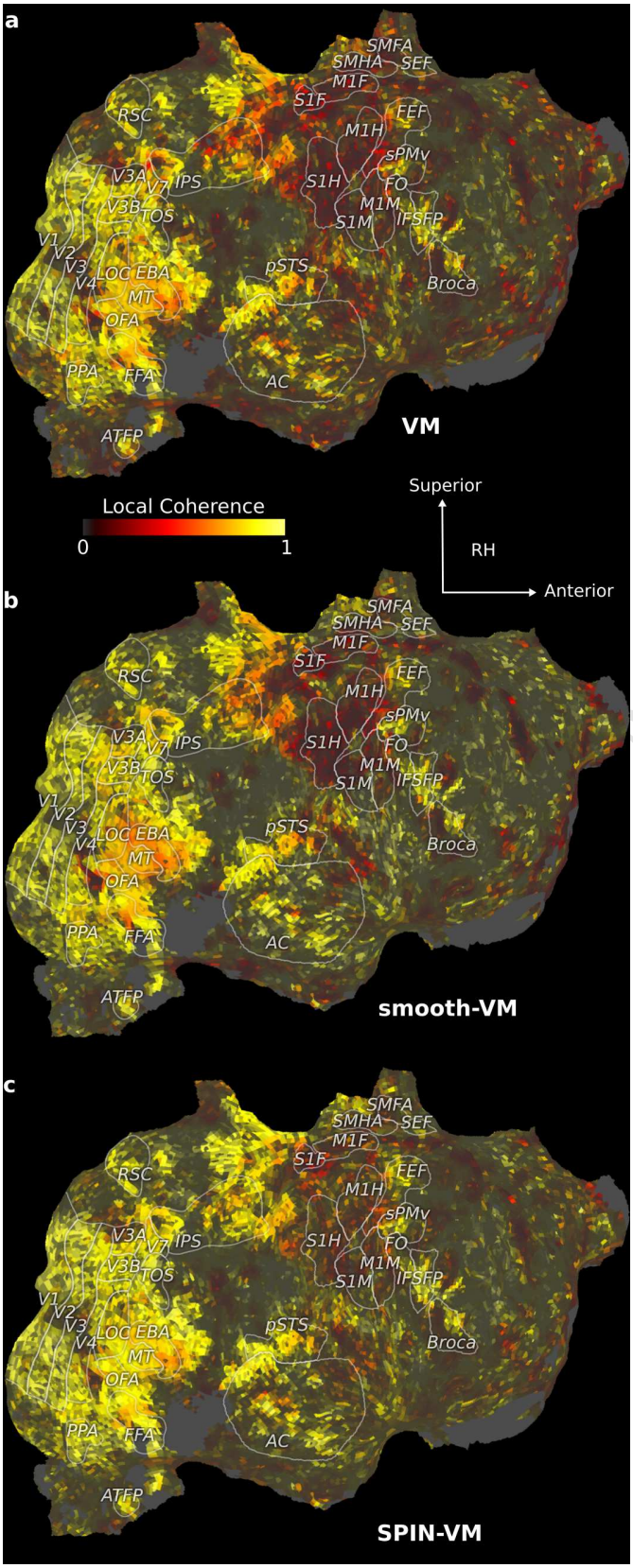


Fig. 8. Cortical flatmaps of local coherence in functional selectivity for the category model. Cortical flatmaps of local coherence for the category model based on model weights estimated by **(a)** VM, **(b)** smooth-VM, and **(c)** SPIN-VM for subject S1. Local coherence was calculated for each voxel based on the standard deviation of model weights across neighboring voxels (see *Local coherence analysis* for details). Yellow indicates higher local coherence compared to red. As expected, spatially smoothing the BOLD responses before implementing VM improves local coherence. However, even though SPIN-VM does not use any spatial smoothing, it yields the most coherent map among all techniques. Improved coherence is observed with SPIN-VM in many voxels distributed across the cortex. The most prominent improvements are observed in high-level visual areas including EBA, MT, and LOC. Formatting is identical to Fig. 4.

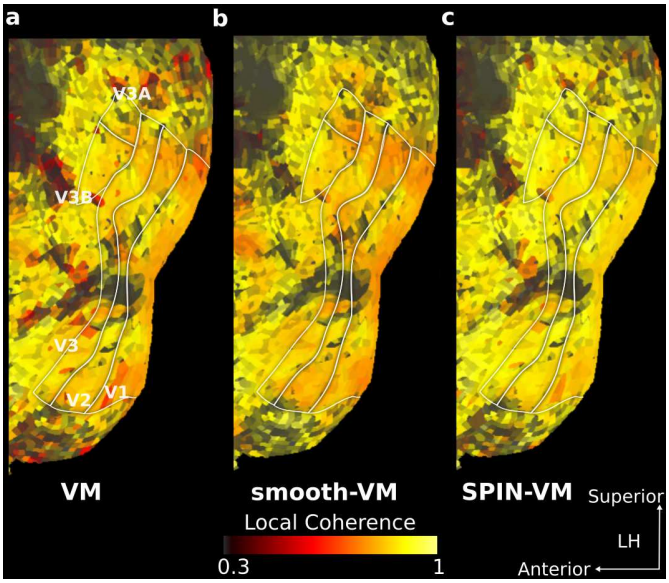


Fig. 9. Cortical flatmaps of local coherence in functional selectivity for the motion-energy model. Cortical flatmaps of local coherence for the motion-energy model based on model weights estimated by **(a)** VM, **(b)** smooth-VM, and **(c)** SPIN-VM for subject S1. Local coherence was calculated for each voxel based on the standard deviation of model weights across neighboring voxels (see *Local coherence analysis* for details). Yellow indicates higher local coherence compared to red. Although SPIN-VM does not use any spatial smoothing, it yields the most coherent map among all techniques. Formatting is identical to Fig. 4.

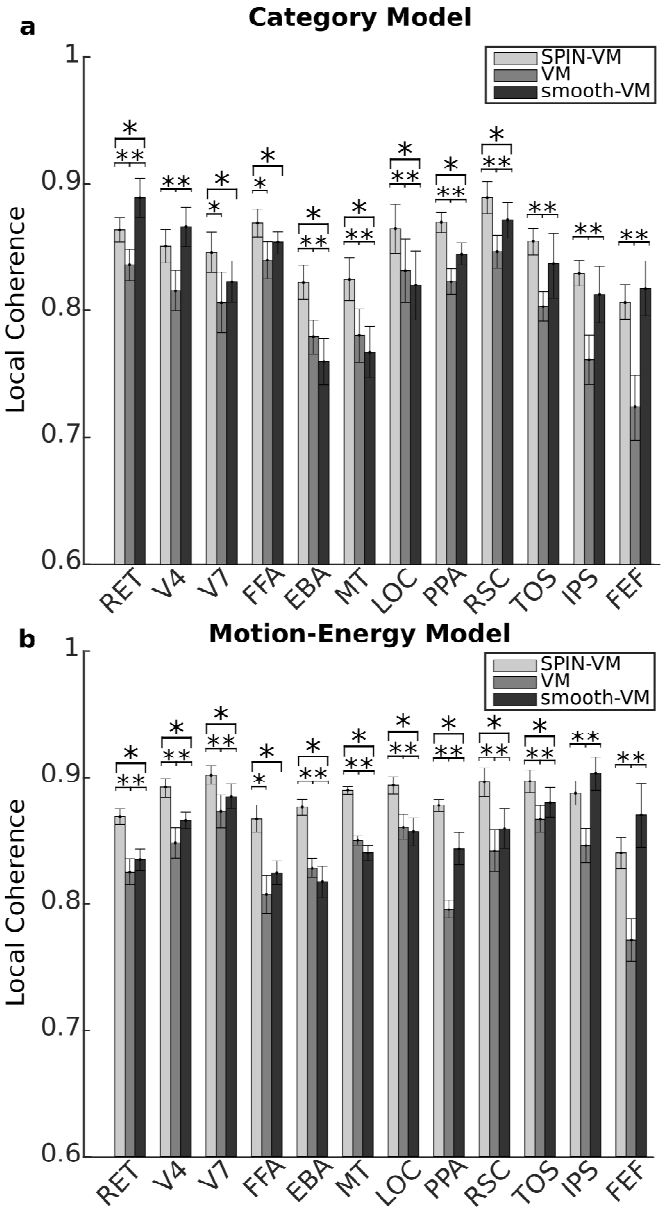


Fig. 10. Local coherence values across functional ROIs. Local coherence values (mean ± SEM) across five subjects in twelve functional ROIs, based on model weights estimated by VM, smooth-VM, and SPIN-VM

for **(a)** the category model and **(b)** the motion-energy model. Brackets indicate significant differences in local coherence ($p < 0.05$, Bootstrap test). (Mean local coherence values for each functional ROI are listed in Supp. Tables 7 and 8 for the category and the motion-energy models, respectively.) For the category model, SPIN-VM results in significantly higher local coherence compared to the other two approaches, especially in lateral occipitotemporal areas including EBA, MT, and LOC ($p < 0.05$), but not in retinotopically organized early visual areas (RET). On the other hand, for the motion-energy model, SPIN-VM results in significantly higher local coherence in retinotopically organized early visual areas (RET and V4), in addition to high-level visual areas in ventral temporal cortex and lateral occipitotemporal cortex including FFA, EBA, MT, LOC, PPA, and RSC ($p < 0.05$).

As expected, SPIN-VM consistently results in significantly higher local coherence than VM in all ROIs for both the category and motion-energy models ($p < 0.05$, Bootstrap test). SPIN-VM also yields significantly higher local coherence than smooth-VM in V7, FFA, EBA, MT, LOC, PPA, and RSC for the category model ($p < 0.05$). No significant difference was observed in V4, TOS, IPS, and FEF ($p > 0.19$). For the motion-energy model, SPIN-VM yields significantly higher local coherence than smooth-VM in all ROIs ($p < 0.05$), except IPS and FEF for which no significant difference was observed ($p > 0.88$). These results confirm that both semantic and Gabor maps produced by SPIN-VM are significantly more coherent compared to those given by VM and smooth-VM.

4. Discussion

Voxelwise modeling (VM) is a powerful framework that can accurately predict single voxel responses evoked by complex natural stimuli, and that can provide an explicit description of how information is represented in individual voxels (Naselaris et al., 2011). However, VM disregards response correlations across neighboring voxels as single-voxel models are fit independently. With high measurement noise, this can diminish sensitivity in assessment of functional selectivity. Here, we proposed a spatially-informed voxelwise modeling (SPIN-VM) technique to address this limitation. SPIN-VM uses regularization across neighboring voxels in addition to regularization across model features. As a result, it improves model performance and yields improved sensitivity in assessment of fine-grained cortical representations.

We optimized the regularization parameters in SPIN-VM across the feature dimension (λ_{feat}) and spatial dimension (λ_{nei}) for each individual voxel separately. In addition, a weighted graph Laplacian is utilized to characterize the extent and distribution of shared information across neighboring voxels. This helps improve sensitivity in detecting functional selectivity of individual voxels. We tested various window sizes and weighting functions to optimize the Laplacian. A Gaussian weighting function with a window size of 3 was observed to yield near-optimal performance broadly across cortical voxels. However, further performance improvements might be possible by optimizing these hyperparameters for each individual voxel separately at the expense of added computational burden.

SPIN-VM has several important advantages over conventional fMRI analyses. Traditional univariate techniques including SPM and functional localizers typically assume smoothness of BOLD responses across contiguous voxels and apply explicit spatial smoothing to increase SNR. This reduces spatial precision as functional selectivity differences across individual voxels are blurred. To increase sensitivity, MVPA was proposed that analyzes the responses of multiple voxels to classify BOLD response patterns into discrete experimental conditions (Haxby, 2012; Norman et al., 2006). While MVPA does not use spatial smoothing, classifier weights are estimated for multiple voxels at once, so they may not accurately reflect the contribution of individual voxels to the represented information. This in turn renders the interpretation of classifier weights difficult (Haufe et al., 2014). In contrast, SPIN-VM utilizes information across neighboring voxels while still optimizing performance for single-voxel response prediction. Thus, SPIN-VM is more powerful in examining fine-grained representations in single voxels compared to both standard univariate and multivariate techniques.

Several methods were previously introduced to leverage shared information across contiguous voxels in order to improve model performance (Grosenick et al., 2013; Katanoda et al., 2002; Penny et al., 2005; Wen and Li, 2016). A joint modeling approach was proposed (Katanoda et al., 2002) that models pooled voxel responses to estimate the weights for the central voxel within a neighborhood. A related approach estimates model weights for voxels within a searchlight separately, and then averages model weights for a given voxel across the multiple distinct searchlights in which it appears (Wen and Li, 2016). In this latter method, the averaging of model weights and prediction scores across searchlights may lead to suboptimal selection of regularization parameters and excessive smoothing of functional selectivity. Moreover, iterative model estimation is performed that can be computationally demanding. Thus, although no spatial smoothing is used, joint-modeling approaches commonly average information contained within the neighborhood during model fitting. This can reduce spatial precision and introduce difficulty in interpreting single-voxel model weights. While SPIN-VM also pools information across a neighborhood, model weights are estimated based on the prediction accuracy of unaveraged single-voxel responses. Therefore, SPIN-VM retains higher sensitivity to functional selectivity in individual voxels.

An alternative approach for utilizing shared information across spatially contiguous voxels is to use spatial

priors (Grosenick et al., 2013; Penny et al., 2005). A previous study proposed a Laplacian operator to penalize differences across model weights of neighboring voxels as in SPIN-VM (Penny et al., 2005). However, in that previous study, no regularization was performed across model features, potentially reducing sensitivity to functional selectivity and limiting utility in analysis of naturalistic fMRI experiments that contain thousands of stimulus features. Another study proposed a graph-constrained operator to implement spatial priors (Grosenick et al., 2013). Graph-constrained operators were demonstrated to improve classification performance for discrete experimental task conditions based on BOLD responses. However, the utility of this approach for fitting encoding models was not considered. Note that both previous methods incorporating spatial priors involve Monte Carlo sampling, so they are computationally more demanding than SPIN-VM.

Response correlations across neighboring voxels can partly be attributed to correlations in stimulus-driven portion of BOLD responses, and partly due to intrinsic noise correlations in BOLD responses (Henriksson et al., 2015). Note that while SPIN-VM utilizes shared information across neighboring voxels, it still aims to fit models that best explain single-voxel BOLD responses in terms of stimulus features. Therefore, if the noise correlations are the dominant factor in driving the response correlations, this will render SPIN-VM less effective in improving model performance. In the natural movie dataset examined here, we observe that SPIN-VM yields higher prediction scores compared to VM across many early- and high-level visual areas, as well as broadly across non-visual cortex. This suggests that a substantial portion of response correlations is stimulus-driven.

In the current study, we find that regularization of model weights across spatial neighborhoods outperforms that based on functional neighborhoods. Because the Laplacian matrix that governs the regularization of model weights is based on inter-voxel distances, this result may be partly attributed to the way that inter-voxel distances are calculated. It is possible that functional distance measurements on inherently noisy BOLD responses might be biased in a way that limits model performance. That said, combining regularization terms across both spatial and functional neighborhoods can potentially be an effective approach to further improve model performance. Our initial empirical observations suggest that a trivial combination of the two approaches—where a spatial neighborhood is selected but voxels within the neighborhood are weighted according to their functional similarity to the central voxel—does not offer any notable improvement (Supp. Fig. 12). However, enhanced performance may be viable by solving a multi-objective optimization problem where both spatial and functional Laplacian matrices are included. This remains an important topic for future investigation.

Here spatial regularization of model weights is performed via L2-norm regularization based on Laplacian matrices of size $(N_{vox}) \times (N_{vox})$, imposing a heavy computational burden. One way to circumvent this would be to truncate the Schur decomposition such that the dimensions of $\mathbf{U}$ and $\mathbf{S}$ corresponding to the smallest eigenvalues are selected (i.e., the lowest frequency components of the Laplacian). A systematic comparison of alternative regularization approaches remains important future work.

In conclusion, we introduced a spatially-informed VM framework that incorporates correlated information across contiguous voxels. Compared to VM, the proposed technique offers improved performance in measuring category and motion-energy selectivity during natural vision. Overall, SPIN-VM yields higher prediction scores in single voxels, increased sensitivity to functional selectivity differences across voxels, and improved utility in assessment of coherent information representations. Therefore, SPIN-VM is a promising tool for analyzing fMRI data collected during naturalistic experiments.

5. Acknowledgments. The authors declare no competing financial interests. We thank A. Vu, N. Bilenko, J. Gao, A. Huth, S. Nishimoto, and J.L. Gallant for assistance in various aspects of this research. The work was supported in part by a National Eye Institute Grant (EY019684), by a Marie Curie Actions Career Integration Grant (PCIG13-GA-2013-618101), by a European Molecular Biology Organization Installation Grant (IG 3028), by a TUBA GEBIP 2015 fellowship, and by a Science Academy BAGEP 2017 award.

6. References

Adler, R.J., Firman, D., 1981. A Non-Gaussian Model for Random Surfaces. Philos. Trans. R. Soc. London. Ser. A Math. Phys. Sci. 303, 433–462.

Connolly, J.D., Goodale, M.A., Desouza, J.F.X., Menon, R.S., Vilis, T., 2000. A Comparison of Frontoparietal fMRI Activation During Anti-Saccades and Anti-Pointing. J. Neurophysiol. 84, 1645–1655. doi:10.1152/jn.2000.84.3.1645

Çukur, T., Huth, A.G., Nishimoto, S., Gallant, J.L., 2016. Functional Subdomains within Scene-Selective Cortex: Parahippocampal Place Area, Retrosplenial Complex, and Occipital Place Area. J. Neurosci. 36, 10257–10273. doi:10.1523/JNEUROSCI.4033-14.2016

Çukur, T., Huth, A.G., Nishimoto, S., Gallant, J.L., 2013a. Functional Subdomains within Human FFA. J. Neurosci. 33, 16748–16766. doi:10.1523/JNEUROSCI.1259-13.2013

Çukur, T., Nishimoto, S., Huth, A.G., Gallant, J.L., 2013b. Attention during natural vision warps semantic representation across the human brain. Nat. Neurosci. 16, 763. doi:10.1038/nn.3381

Dale, A.M., Fischl, B., Sereno, M.I., 1999. Cortical Surface-Based Analysis I. Segmentation and Surface Reconstruction. Neuroimage 9, 179–194. doi:10.1006/nimg.1998.0395

David, S. V., Gallant, J.L., 2005. Predicting neuronal responses during natural vision. Netw. Comput. Neural Syst. 16, 239–260. doi:10.1080/09548980500464030

Downing, P., Jiang, Y., Shuman, M., Kanwisher, N., 2001. A Cortical Area Selective for Visual Processing of the Human Body. Science 293, 2470–2473. doi:10.1126/science.1063414

Dumoulin, S.O., Wandell, B.A., 2008. Population receptive field estimates in human visual cortex. Neuroimage 39, 647–660. doi:10.1016/j.neuroimage.2007.09.034

Engel, S.A., Glover, G.H., Wandell, B.A., 1997. Retinotopic organization in human visual cortex and the spatial precision of functional MRI. Cereb. Cortex 7, 181–192. doi:10.1093/cercor/7.2.181

Erwin, E., Obermayer, K., Schulten, K., 1995. Models of Orientation and Ocular Dominance Columns in the Visual Cortex: A Critical Comparison. Neural Comput. 7, 425–468. doi:10.1162/neco.1995.7.3.425

Etzel, J.A., Zacks, J.M., Braver, T.S., 2013. Searchlight analysis: Promise, pitfalls, and potential. Neuroimage 78, 261–269. doi:10.1016/j.neuroimage.2013.03.041

Friston, K.J., Holmes, A.P., Worsley, K.J., Poline, J.-P., Frith, C.D., Frackowiak, R.S.J., 1994. Statistical Parametric Maps in Functional Imaging: A General Linear Approach. Hum. Brain Mapp. 2, 189–210. doi:10.1002/hbm.460020402

Gao, J.S., Huth, A.G., Lescroart, M.D., Gallant, J.L., 2015. Pycortex: an interactive surface visualizer for fMRI. Front. Neuroinform. 9, 23. doi:10.3389/fninf.2015.00023

Grefkes, C., Fink, G.R., 2005. The functional organization of the intraparietal sulcus in humans and monkeys. J. Anat. 207, 3–17. doi:10.1111/j.1469-7580.2005.00426.x

Greve, D.N., Fischl, B., 2009. Accurate and robust brain image alignment using boundary-based registration. Neuroimage 48, 63–72. doi:10.1016/j.neuroimage.2009.06.060

Grosenick, L., Klingenberg, B., Katovich, K., Knutson, B., Taylor, J.E., 2013. Interpretable whole-brain prediction analysis with GraphNet. Neuroimage 72, 304–321. doi:10.1016/j.neuroimage.2012.12.062

Hansen, K.A., Kay, K.N., Gallant, J.L., 2007. Topographic Organization in and near Human Visual Area V4. J. Neurosci. 27, 11896–11911. doi:10.1523/JNEUROSCI.2991-07.2007

Haufe, S., Meinecke, F., Görgen, K., Dähne, S., Haynes, J.D., Blankertz, B., Bießmann, F., 2014. On the interpretation of weight vectors of linear models in multivariate neuroimaging. Neuroimage 87, 96–110. doi:10.1016/j.neuroimage.2013.10.067

Haxby, J. V., 2012. Multivariate pattern analysis of fMRI: The early beginnings. Neuroimage 62, 852–855. doi:10.1016/j.neuroimage.2012.03.016

Henriksson, L., Khaligh-Razavi, S.M., Kay, K., Kriegeskorte, N., 2015. Visual representations are dominated by intrinsic fluctuations correlated between areas. Neuroimage 114, 275–286. doi:10.1016/j.neuroimage.2015.04.026

Huth, A.G., de Heer, W.A., Griffiths, T.L., Theunissen, F.E., Gallant, J.L., 2016. Natural speech reveals the semantic maps that tile human cerebral cortex. Nature 532, 453–458. doi:10.1038/nature17637

Huth, A.G., Nishimoto, S., Vu, A.T., Gallant, J.L., 2012. A Continuous Semantic Space Describes the Representation of Thousands of Object and Action Categories across the Human Brain. Neuron 76, 1210–1224. doi:10.1016/j.neuron.2012.10.014

Jenkinson, M., Bannister, P., Brady, M., Smith, S., 2002. Improved optimization for the robust and accurate linear registration and motion correction of brain images. Neuroimage 17, 825–841. doi:10.1016/S1053-8119(02)91132-8

Kamitani, Y., Sawahata, Y., 2010. Spatial smoothing hurts localization but not information: Pitfalls for brain mappers. Neuroimage 49, 1949–1952. doi:10.1016/j.neuroimage.2009.06.040

Kanwisher, N., McDermott, J., Chun, M.M., 1997. The Fusiform Face Area: A Module in Human Extrastriate Cortex Specialized for Face Perception. J. Neurosci. 17, 4302–4311. doi:10.3410/f.717989828.793472998

Katanoda, K., Matsuda, Y., Sugishita, M., 2002. A Spatio-temporal Regression Model for the Analysis of Functional MRI Data. Neuroimage 17, 1415–1428. doi:10.1006/nimg.2002.1209

Kay, K.N., Naselaris, T., Prenger, R.J., Gallant, J.L., 2008. Identifying natural images from human brain activity. Nature 452, 352. doi:10.1038/nature06713

Kriegeskorte, N., Bandettini, P., 2007. Analyzing for information, not activation, to exploit high-resolution fMRI. Neuroimage 38, 649–662. doi:10.1016/j.neuroimage.2007.02.022

Kriegeskorte, N., Goebel, R., Bandettini, P., 2006. Information-based functional brain mapping. Proc. Natl. Acad. Sci. U. S. A. 103, 3863–3868. doi:10.1073/pnas.0600244103

Lee, D., Port, N.L., Kruse, W., Georgopoulos, A.P., 1998. Variability and Correlated Noise in the Discharge of Neurons in Motor and Parietal Areas of the Primate Cortex. J. Neurosci. 18, 1161–1170.

Lescroart, M.D., Stansbury, D.E., Gallant, J.L., 2015. Fourier power, subjective distance, and object categories all provide plausible models of BOLD responses in scene-selective visual areas. Front. Comput. Neurosci. 9, 135. doi:10.3389/fncom.2015.00135

Miller, G., 1995. WordNet: A Lexical Database for English. Commun. ACM 38, 39–41. doi:10.1145/219717.219748

Mitchell, T.M., Shinkareva, S. V., Carlson, A., Chang, K.-M., Malave, V.L., Mason, R.A., Just, M.A., 2008. Predicting Human Brain Activity Associated with the Meanings of Nouns. Science 320, 1191–1195. doi:10.1126/science.1152876

Naselaris, T., Kay, K.N., Nishimoto, S., Gallant, J.L., 2011. Encoding and decoding in fMRI. Neuroimage 56, 400–410. doi:10.1016/j.neuroimage.2010.07.073

Naselaris, T., Prenger, R.J., Kay, K.N., Oliver, M., Gallant, J.L., 2009. Bayesian Reconstruction of Natural Images from Human Brain Activity. Neuron 63, 902–915. doi:10.1016/j.neuron.2009.09.006

Nelder, J.A., Wedderburn, R.W.M., 1972. Generalized Linear Models. J. R. Stat. Soc. 135, 370–384.

Nishimoto, S., Vu, A.T., Naselaris, T., Benjamini, Y., Yu, B., Gallant, J.L., 2011. Reconstructing Visual Experiences from Brain Activity Evoked by Natural Movies. Curr. Biol. 21, 1641–1646. doi:10.1016/j.cub.2011.08.031

Norman, K.A., Polyn, S.M., Detre, G.J., Haxby, J. V, 2006. Beyond mind-reading: multi-voxel pattern analysis of fMRI data. Trends Cogn. Sci. 10, 424–430. doi:10.1016/j.tics.2006.07.005

Penny, W.D., Trujillo-Barreto, N.J., Friston, K.J., 2005. Bayesian fMRI time series analysis with spatial priors. Neuroimage 24, 350–362. doi:10.1016/j.neuroimage.2004.08.034

Pouget, A., Dayan, P., Zemel, R., 2000. Information Processing with Population Codes. Nat. Rev. Neurosci. 1, 125. doi:10.1038/35039062

Schneidman, E., Berry, M.J., Segev, R., Bialek, W., 2006. Weak pairwise correlations imply strongly correlated network states in a neural population. Nature 440, 1007. doi:10.1038/nature04701

Serences, J.T., Saproo, S., 2012. Computational advances towards linking BOLD and behavior. Neuropsychologia 50, 435–446. doi:10.1016/j.neuropsychologia.2011.07.013

Smith, M.A., Kohn, A., 2008. Spatial and Temporal Scales of Neuronal Correlation in Primary Visual Cortex. J. Neurosci. 28, 12591–12603. doi:10.1523/JNEUROSCI.2929-08.2008

Smith, S.M., 2002. Fast robust automated brain extraction. Hum. Brain Mapp. 17, 143–155. doi:10.1002/hbm.10062

Spiridon, M., Fischl, B., Kanwisher, N., 2006. Location and Spatial Profile of Category-Specific Regions in Human Extrastriate Cortex. Hum. Brain Mapp. 27, 77–89. doi:10.1002/hbm.20169

Sprague, T., Serences, J., 2013. Attention modulates spatial priority maps in the human occipital, parietal and frontal cortices. Nat. Neurosci. 16, 1879. doi:10.1038/nn.3574

Srinivasan, R., Golomb, J.D., Martinez, A.M., 2016. A Neural Basis of Facial Action Recognition in Humans. J. Neurosci. 36, 4434–4442. doi:10.1523/JNEUROSCI.1704-15.2016

Subbian, K., Banerjee, A., 2013. Climate Multi-model Regression Using Spatial Smoothing, in: Proceedings of the 2013 SIAM International Conference on Data Mining. Society for Industrial and Applied Mathematics, pp. 324–332. doi:10.1137/1.9781611972832.36

Thirion, B., Duchesnay, E., Hubbard, E., Dubois, J., Poline, J.B., Lebihan, D., Dehaene, S., 2006. Inverse retinotopy: Inferring the visual content of images from brain activation patterns. Neuroimage 33, 1104–1116. doi:10.1016/j.neuroimage.2006.06.062

Tootell, R., Hadjikhani, N.K., Vanduffel, W., Liu, A.K., Mendola, J.D., Sereno, M.I., Dale, A.M., 1998. Functional analysis of primary visual cortex (V1) in humans. Proc. Natl. Acad. Sci. U. S. A. 95, 811–817.

Tootell, R.B.H., Reppas, J.B., Kwong, K.K., Malach, R., Born, R.T., Brady, T.J., Rosen, B.R., Belliveaul, J.W., 1995. Functional Analysis of Human MT and Related Visual Cortical Areas Using Magnetic Resonance Imaging. J. Neurosci. 15, 3215–3230.

Walbrin, J., Downing, P., Koldewyn, K., 2018. Neural responses to visually observed social interactions. Neuropsychologia 112, 31–39. doi:10.1016/j.neuropsychologia.2018.02.023

Wen, Z., Li, Y., 2016. A spatial□constrained multi□target regression model for human brain activity prediction. Appl. Informatics 3, 10. doi:10.1186/s40535□016□0026□x

Zarahn, E., Aguirre, G.K., D'esposito, M., 1997. Empirical Analyses of BOLD fMRI Statistics. Neuroimage 5, 179–197. doi:10.1006/nimg.1997.0263

Highlights

- A novel spatially informed voxelwise modeling (SPIN-VM) technique is proposed.
- Correlations across neighboring voxels are leveraged during estimation of functional selectivity.
- Compared to VM, SPIN-VM offers improved accuracy in predicting single-voxel BOLD responses.
- SPIN-VM is more sensitive in revealing coherent information representations across cortex.
- SPIN-VM is a powerful method for modeling fMRI data from naturalistic experiments.